Supplement to “Two Applications of Wild Bootstrap Methods to Improve Inference in Cluster-IV Models”

Keith Finlay[*] Leandro M. Magnusson [†]

U.S. Census Bureau University of Western Australia

March 6, 2019

[*]U.S. Census Bureau, Room 5K132D, 4600 Silver Hill Road, Washington, DC 20233, USA, kfinlay@gmail.com.

[†]Department of Economics, 35 Stirling Highway M251, Business School, University of Western Australia, Crawley, WA 6009, Australia, phone: +61 (8) 6488-2924, fax: +61 (8) 6488-1016, leandro.magnusson@uwa.edu.au (corresponding author).

Contents

S.1 Further results from the Miguel et al. (2004) application

We compute the 95% projection-based confidence intervals for both endogenous variables. The projection-based confidence interval for θ_1 is formed by all points of the 95% confidence set plane of θ_1-θ_2 projected into space θ_1. We also compute ratios of the asymptotic and wild bootstrap AR intervals relative to the asymptotic Wald interval along with the first-stage robust F-test for each parameter. The results are reported in Table S.1. As in the other figures from this application, many of the confidence intervals are unbounded.

The lower panel of Table S.1 shows p-values from the joint null hypotheses that the primary parameters are both equal to zero. As opposed to the p-values from the the Acemoglu et al. (2012) application, the p-values based on the asymptotic Wald tests are larger than the p-values obtained from the asymptotic and bootstrapped AR tests. Only the last specification indicates a statistically significant effect of economic performance on civil war at the 1% significance level. From the projected confidence intervals, however, we cannot conclude if this effect is negative or positive.

[Table 1 about here.]

S.2 Tests robust to the presence of weak instruments

The AR, KLM, and CLR tests were originally developed under the assumption that the distribution of the errors is iid, but they have been adapted to allow for arbitrary heteroskedasticity or cluster dependence of the residuals (Chernozhukov and Hansen, 2008; Finlay and Magnusson, 2009). The equations in (7) can be further rewritten as

$$\begin{cases} \hat{\delta}_w(\theta_0) = & \underbrace{\Pi_w \mathrm{d}(\theta_0) + \mathrm{H}\gamma}_{\delta_w(\theta_0)} + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{e}(\theta_0) \\ \widehat{\Pi}_w = & \Pi_w + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}, \end{cases} \tag{CS 1}$$

where $\hat{\delta}_w(\theta_0) = [\hat{\delta}_z(\theta_0)', \hat{\delta}_x(\theta_0)']' = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}(\theta_0)$ and $\widehat{\Pi}_w = [\widehat{\Pi}_z', \widehat{\Pi}_x']' = (\mathbf{W}'\mathbf{W})^{-1}$ $\mathbf{W}'\mathbf{y}_2$ are, respectively, the OLS estimators of the reduced-form parameters $\delta_w(\theta_0)$ and Π_w. The $k_w(p+1) \times k_w(p+1)$ "sandwich" matrix that corresponds to the cluster-robust estimator of the variance of $[\hat{\delta}_w(\theta_0)', \mathrm{vec}(\widehat{\Pi}_w)']$ is

$$\widehat{\Omega}(\theta_0) = \left(\mathrm{I}_{p+1} \otimes \mathbf{W}'\mathbf{W}\right)^{-1} \widehat{\Xi}(\theta_0) \left(\mathrm{I}_{p+1} \otimes \mathbf{W}'\mathbf{W}\right)^{-1}, \tag{CS 2}$$

where $\widehat{\Xi}(\theta_0) = \left[\widehat{\Xi}_{ee}(\theta_0), \widehat{\Xi}_{eV}(\theta_0) : \widehat{\Xi}_{Ve}(\theta_0), \widehat{\Xi}_{VV}(\theta_0)\right]$ is the estimator of the $k_w(p+1) \times k_w(p+1)$ variance matrix of $\mathrm{vec}(\mathbf{W}'\mathbf{e}(\theta_0), \mathbf{W}'\mathbf{V})$.[1] Let us introduce two more statistics before presenting the cluster versions of the KLM and CLR tests:

$$\tilde{\lambda}_{\mathrm{KLM}}(\theta_0) = \widetilde{\Pi}_z(\theta_0)' \left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1} \hat{\delta}_z(\theta_0) \text{ and}$$
$$\widetilde{\Pi}_z(\theta_0) = \mathrm{mat}\left(\hat{\pi}_z - \widehat{\Omega}_{\pi_z\delta_z}(\theta_0)\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\hat{\delta}_z(\theta_0)\right), \tag{CS 3}$$

where $\tilde{\lambda}_{\mathrm{KLM}}(\theta_0)$ is the Lagrange multiplier of a constrained minimum-distance minimization problem,[2] $\hat{\pi}_z = \mathrm{vec}(\widehat{\Pi}_z)$ is a $k_z p \times 1$ vector, and mat is the rematricizing operator that maps the $k_z p \times 1$ vector $\mathrm{vec}(\widehat{\Pi}_z)$ onto the $k_z \times p$ matrix $\widehat{\Pi}_z$. The submatrices of $\widehat{\Omega}(\theta_0)$, $\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)$ and $\widehat{\Omega}_{\pi_z\delta_z}(\theta_0)$, are the variance and covariance estimators of $\hat{\delta}_z(\theta_0)$ and $\left(\hat{\delta}_z(\theta_0)', \hat{\pi}_z'\right)'$, respectively. The estimator of the variance of $\tilde{\lambda}_{\mathrm{KLM}}(\theta_0)$ is $\widehat{\mathrm{Var}}\left(\tilde{\lambda}_{\mathrm{KLM}}(\theta_0)\right) = \widetilde{\Pi}_z(\theta_0)'\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\widetilde{\Pi}_z(\theta_0)$. The tests for the structural parameter that are robust to the presence of weak instruments for the cluster IV model are defined as follows:

Definition 1 (Weak instrument robust tests with clustered residuals)**.** The AR, KLM, and CLR statistics for testing the null hypothesis $H_0 : \mathrm{d}(\theta_0) = 0$ are, respectively:

$$\begin{aligned}
\Lambda_{\mathrm{AR}}(\theta_0) &\equiv \hat{\delta}_z(\theta_0)'\left[\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)\right]^{-1}\hat{\delta}_z(\theta_0) \xrightarrow{d} \chi^2(k_z), \\
\Lambda_{\mathrm{KLM}}(\theta_0) &\equiv \tilde{\lambda}_{\mathrm{KLM}}(\theta_0)'\left[\widehat{\mathrm{Var}}\left(\tilde{\lambda}_{\mathrm{KLM}}(\theta_0)\right)\right]^{-1}\tilde{\lambda}_{\mathrm{KLM}}(\theta_0) \xrightarrow{d} \chi^2(p), \text{ and} \\
\Lambda_{\mathrm{CLR}}(\theta_0) &\equiv \left\{\frac{1}{2}\Lambda_{\mathrm{AR}}(\theta_0) - \mathrm{rk}(\theta_0) + \right. \\
&\qquad \left. \sqrt{\left[\Lambda_{\mathrm{AR}}(\theta_0) + \mathrm{rk}(\theta_0)\right]^2 - 4\left[\Lambda_{\mathrm{AR}}(\theta_0) - \Lambda_{\mathrm{KLM}}(\theta_0)\right] \times \mathrm{rk}(\theta_0)}\right\},
\end{aligned}$$

where $\mathrm{rk}(\theta_0)$ is a statistic for testing the rank of $\widetilde{\Pi}_z(\theta_0)$. The symbol "$\xrightarrow{d}$"stands for convergence in distribution as $G \to +\infty$, and $\chi^2(s)$ is the chi-squared distribution with s degrees of freedom. The CLR-statistic converges to a nonpivotal distribution. However, its critical values, for a given value of $\mathrm{rk}(\theta_0)$, can be simulated from independent $\chi^2(k_z - p)$ and $\chi^2(p)$ distributions.

The above tests have the correct sizes asymptotically even when the structural parameter θ is not identified. When $\Pi_z \approx 0$, the tests will not have power to reject the null hypothesis, indicating the presence of weak instruments. The degrees of freedom of the distribution of the asymptotic AR test depends on k_z, the number of excluded instruments, which can be larger

[1] Details about the computation of $\widehat{\Xi}(\theta_0)$ are in Section S.4.
[2] See the derivation in Section S.3.

than p, the number of tested parameters. The larger the difference $k_z - p$, the less powerful the AR test is. The number of degrees of freedom for the KLM test is equal to the number of tested structural parameters, independent of the number of excluded instruments. Nevertheless, the KLM test, as an LM type of test, loses power at the local extrema and inflection points of the AR test. The CLR test, because it is a function of the AR test, does not exhibit the spurious decline of power experienced by the KLM test.

Remark S.2.1. If θ is scalar, the rank statistic rk is defined as:

$$\mathrm{rk}\left(\theta_0\right) \equiv \widetilde{\Pi}_z\left(\theta_0\right)' \left[\widehat{\mathrm{Var}}\left(\widetilde{\Pi}_z\left(\theta_0\right)\right)\right]^{-1} \widetilde{\Pi}_z\left(\theta_0\right),$$

where $\widehat{\mathrm{Var}}\left(\widetilde{\Pi}_z\left(\theta_0\right)\right) = \widehat{\Omega}_{\pi_z\pi_z} - \widehat{\Omega}_{\pi_z\delta_z}\left(\theta_0\right)\left[\widehat{\Omega}_{\delta_z\delta_z}\left(\theta_0\right)\right]^{-1}\widehat{\Omega}_{\delta_z\pi_z}\left(\theta_0\right)$. If θ is not scalar, then the rank statistics proposed by Robin and Smith (2000) using the moment-variance weighting described in Andrews and Guggenberger (2017) should be used.[3]

S.3 Derivation of the Lagrange multiplier estimator

Our KLM test is the Lagrange multiplier test derived from the following restricted minimization problem under $H_0 : \mathrm{d}\left(\theta_0\right) = 0$:

$$\begin{array}{l} \min\limits_{\pi_w,\,\gamma} \\ \text{s.t. } \mathrm{d}\left(\theta_0\right) = 0 \end{array} \quad \frac{1}{2}\begin{pmatrix} \hat{\delta}_w\left(\theta_0\right) - \Pi_w \mathrm{d}\left(\theta_0\right) - \mathrm{H}\gamma \\ \hat{\pi}_w - \pi_w \end{pmatrix}' \left[\widehat{\Omega}\left(\theta_0\right)\right]^{-1} \begin{pmatrix} \hat{\delta}_w\left(\theta_0\right) - \Pi_w \mathrm{d}\left(\theta_0\right) - \mathrm{H}\gamma \\ \hat{\pi}_w - \pi_w \end{pmatrix}, \qquad \text{(CS 4)}$$

where $\mathrm{H}= \left[\begin{array}{cc} 0' & \mathrm{I}_{k_x}' \end{array}\right]'$ and the estimator $\widehat{\Omega}\left(\theta_0\right)$ is defined as in equation (CS 2). In the following, we omit (θ_0) from $\delta_w\left(\theta_0\right)$, $\mathrm{d}(\theta_0)$, and $\widehat{\Omega}\left(\theta_0\right)$ to simplify the exposition. From the first-order condition, we obtain:

$$\mathrm{H}'\left(\widehat{\Omega}_{\delta_w\delta_w}\right)^{-1}\left(\hat{\delta}_w - \mathrm{H}\tilde{\gamma}\left(\theta_0\right)\right) = 0, \text{ and } \widetilde{\Pi}_w\left(\theta_0\right)'\left(\widehat{\Omega}_{\delta_w\delta_w}\right)^{-1}\left(\hat{\delta}_w - \mathrm{H}\tilde{\gamma}\left(\theta_0\right)\right) = \tilde{\lambda}\left(\theta_0\right), \qquad \text{(CS 5)}$$

from which we derive $\tilde{\gamma}\left(\theta_0\right) = [\mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\mathrm{H}]^{-1}\mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\hat{\delta}_w$ and $\tilde{\lambda}\left(\theta_0\right) = \widetilde{\Pi}_w\left(\theta_0\right)'$ $(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\,\mathrm{M}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}\hat{\delta}_w$, where $\mathrm{M}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}} = \mathrm{I}-\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}$ and $\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}=\mathrm{H}\,[\mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\mathrm{H}]^{-1}\,\mathrm{H}'\,(\widehat{\Omega}_{\delta_w\delta_w})^{-1}$ is an oblique projection.[4] Using the fact that $(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\mathrm{M}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}} =$

[3] Andrews and Guggenberger (2017) show that rank tests based on the variance of the Jacobian, as in Kleibergen and Paap (2006), do not necessarily have the correct asymptotic size in models with two or more endogenous explanatory variables.

[4] The oblique projection $\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}$ satisfies the properties $\mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}} = \mathrm{H}'(\widehat{\Omega}_{\delta_w\delta_w})^{-1}$ and $\mathrm{P}_{\mathrm{H}}^{(\widehat{\Omega}_{\delta_w\delta_w})^{-1}}\mathrm{H}=\mathrm{H}$.

$[(\widehat{\Omega}_{\delta_z\delta_z})^{-1},\ 0:\ 0\ ,0]$, further simplification allows us to write the Lagrange multiplier estimator as $\tilde{\lambda}(\theta_0) = \widetilde{\Pi}_z(\theta_0)'\ (\widehat{\Omega}_{\delta_z\delta_z})^{-1}\ \hat{\delta}_z$. An estimator of the variance of $\tilde{\lambda}(\theta_0)$ conditional on $\widetilde{\Pi}_z(\theta_0)$ is $\widehat{\mathrm{Var}}(\tilde{\lambda}(\theta_0)) = \widetilde{\Pi}_z(\theta_0)'(\widehat{\Omega}_{\delta_z\delta_z})^{-1}\ \widetilde{\Pi}_z(\theta_0)$. Finally, the estimator for Π_w derived from equation (CS 5) is

$$\tilde{\pi}_w(\theta_0) = \mathrm{vec}\left(\widehat{\Pi}_w\right) - \widehat{\Omega}_{\pi_w\delta_z}\left(\widehat{\Omega}_{\delta_z\delta_z}\right)^{-1}\hat{\delta}_z,$$

with the variance estimator $\widehat{\mathrm{Var}}(\tilde{\pi}(\theta_0)) = \widehat{\Omega}_{\pi_w\pi_w} - \widehat{\Omega}_{\pi_w\delta_z}\left(\widehat{\Omega}_{\delta_z\delta_z}\right)^{-1}\widehat{\Omega}_{\delta_z\pi_w}$.

Remark S.3.1. The *KLM* test is originally derived from the continuously updating estimator (CUE) objective function rather than from the two-step minimum-distance estimator objective function above, which is a simpler problem to be solved. Because the regression model is linear, the minimum-distance estimates of the untested well-identified parameters are the same as the GMM-CUE estimator under the null assumption (Goldberger and Olkin, 1971).

The AR test also has a Lagrange-multiplier interpretation. It can be rewritten as

$$\Lambda_{\mathrm{AR}}(\theta_0) \equiv \tilde{\lambda}_{\mathrm{AR}}(\theta_0)'\,[\widehat{\mathrm{Var}}(\tilde{\lambda}_{\mathrm{AR}}(\theta_0))]^{-1}\tilde{\lambda}_{\mathrm{AR}}(\theta_0),$$

where $\tilde{\lambda}_{\mathrm{AR}}(\theta_0) = [\widehat{\Omega}_{\delta_z\delta_z}(\theta_0)]^{-1}\hat{\delta}_z(\theta_0)$ is the Lagrange multiplier of the constrained-minimization problem

$$\begin{aligned} \min_{\delta_w} \quad & \frac{1}{2}\left(\hat{\delta}_w - \delta_w\right)'\left[\widehat{\Omega}_{\delta_w\delta_w}\right]^{-1}\left(\hat{\delta}_w - \delta_w\right), \\ \text{s.t. } & \delta_z = 0 \end{aligned} \tag{CS 6}$$

and $\widehat{\mathrm{Var}}\left(\tilde{\lambda}_{\mathrm{AR}}(\theta_0)\right)$ is the estimated variance of $\tilde{\lambda}_{\mathrm{AR}}(\theta_0)$. From the first-order conditions with respect to δ_x and the Lagrange multiplier λ, we find $\tilde{\lambda}_{\mathrm{AR}}(\theta_0) = (\widehat{\Omega}_{\delta_z\delta_z})^{-1}\hat{\delta}_z$, and $\tilde{\delta}_x(\theta_0) = \hat{\delta}_x - \widehat{\Omega}_{\delta_x\delta_z}\tilde{\lambda}_{\mathrm{AR}}(\theta_0)$, which is the same as the minimum-distance estimator for γ in Equation (CS 5).

S.4 Cluster variance-covariance matrix

Define $\mathbf{h}_g(\theta_0) = \sum_{i=1}^{n_g}\mathbf{h}_{i,g}(\theta_0) = \mathbf{w}_g'\mathbf{e}_g(\theta_0)$ with $\mathbf{h}_{i,g} = \mathbf{w}_{i,g}'e_{i,g}(\theta_0)$ and $\mathbf{q}_g = \sum_{i=1}^{n_g}\mathbf{q}_{i,g} = (\mathrm{I}_p \otimes \mathbf{w}_g')\mathbf{v}_g$ with $\mathbf{q}_{i,g} = (\mathrm{I}_p \otimes \mathbf{w}_{i,g}')v_{i,g}$, where $\mathbf{w}_g = (\mathbf{z}_g : \mathbf{x}_g)$, $\mathbf{v}_g = \mathrm{vec}(\mathbf{y}_{2,g} - \mathbf{w}_g\Pi_w)$, $\mathbf{e}_g = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\delta_w$. Let us partition $\Xi(\theta_0)$, the variance matrix in Equation (CS 2), as $\Xi(\theta_0) =$

$[\Xi_{\mathbf{hh}}(\theta_0), \Xi_{\mathbf{hq}}(\theta_0) : \Xi_{\mathbf{qh}}(\theta_0), \Xi_{\mathbf{qq}}(\theta_0)]$. Each component of $\Xi(\theta_0)$ is estimated as

$$\Xi_{\mathbf{rs}}(\theta_0) = \frac{1}{n}\sum_{g=1}^{G}\left(\mathbf{r}_g(\theta_0) - n_g\bar{\mathbf{r}}(\theta_0)\right)\left(\mathbf{s}_g(\theta_0) - n_g\bar{\mathbf{s}}_g(\theta_0)\right)',$$

where $\mathbf{r}_g(\theta_0) = \mathbf{h}_g(\theta_0)$ or $\mathbf{q}_g(\theta_0)$ with $\bar{\mathbf{r}}(\theta_0) = \frac{1}{n}\sum_{g=1}^{G}\mathbf{r}_g(\theta_0)$ and $\mathbf{s}_g(\theta_0) = \mathbf{h}_g(\theta_0)$ or $\mathbf{q}_g(\theta_0)$ with $\bar{\mathbf{s}}(\theta_0) = \frac{1}{n}\sum_{g=1}^{G}\mathbf{s}_g(\theta_0)$. To compute the variance of the weak-instrument-robust tests in Section 4, we replace $(\mathbf{e}_g(\theta_0), \mathbf{v}_g)$ with $(\acute{\mathbf{e}}_g(\theta_0), \hat{\mathbf{v}}_g) = (\mathbf{Y}_g(\theta_0) - \mathbf{w}_g\acute{\delta}_w(\theta_0), \mathbf{y}_{2,g} - \mathbf{w}_g\widehat{\Pi}_w)$. In the case of the EE bootstrap, we use

$$\tilde{\mathbf{h}}_g^*(\theta_0) = \omega_g\tilde{\mathbf{h}}_g^c(\theta_0) = \omega_g\left(\mathbf{w}_g'\tilde{\mathbf{e}}_g(\theta_0) - \frac{n_g}{n}\sum_{g=1}^{G}\mathbf{w}_g'\tilde{\mathbf{e}}_g(\theta_0)\right)$$

in place of $\mathbf{h}_g(\theta_0)$. The remaining variance terms of the EE bootstrap are not computed because we are conditioning on $\widetilde{\Pi}_z(\theta_0)$. For the residual bootstrap cases, the covariance matrix is obtained by substituting $(\mathbf{e}_g(\theta_0), \mathbf{v}_g)$ with $(\mathbf{e}^*_{\mathrm{b},,g}(\theta_0), \mathbf{v}^*_{\mathrm{b},,g})$ in the $\mathbf{h}_g(\theta_0)$ and $\mathbf{q}_g$ equations defined above, where $\mathbf{e}^*_{\mathrm{b},g} = \mathbf{Y}^*_g(\theta_0) - \mathbf{w}_g\delta^*_w(\theta_0)$, $\mathbf{v}^*_{\mathrm{b},g} = \mathbf{y}^*_{2,g}(\theta_0) - \mathbf{w}_g\Pi^*_w(\theta_0)$, and $(\delta^*_w(\theta_0), \Pi^*_w(\theta_0))$ are the bootstrap estimate values.

S.5 Other bootstrap methods

We present two additional residual bootstrap methods for the AR and KLM tests. The first is the multi-equation (ME) residual bootstrap in which bootstrap samples are obtained from drawing the first- and second-stage residuals simultaneously. The second is the Davidson and MacKinnon (2010) bootstrap method for linear IV models adapted for the cluster framework, which we refer to as the DM bootstrap. We also present three bootstrap methods for the Wald test at the end of this Section.

S.5.1 Estimating equations (score) bootstrap for KLM and CLR tests

We can extend the procedure above to compute the KLM and CLR bootstrap tests conditional on $\widetilde{\Pi}_z(\theta_0)$ and $\mathrm{rk}(\theta_0)$. We define the bootstrap estimator of $\tilde{\lambda}_{\mathrm{KLM}}(\theta_0)$ and its variance as:

$$\begin{aligned}\tilde{\lambda}^*_{\mathrm{KLM}}(\theta_0) &= \widetilde{\Pi}_z(\theta_0)'\left[\widetilde{\Omega}^*_{\delta_z\delta_z}(\theta_0)\right]^{-1}\tilde{\delta}^*_z(\theta_0), \text{ and} \\ \widehat{\mathrm{Var}}\left(\tilde{\lambda}^*_{\mathrm{KLM}}(\theta_0)\right) &= \widetilde{\Pi}_z(\theta_0)'\left[\widetilde{\Omega}^*_{\delta_z\delta_z}(\theta_0)\right]^{-1}\widetilde{\Pi}_z(\theta_0).\end{aligned} \tag{CS 7}$$

The bootstrapped KLM test is obtained by replacing $\tilde{\lambda}^*_{\text{KLM}}(\theta_0)$ and $\widehat{\text{Var}}\left(\tilde{\lambda}^*_{\text{KLM}}(\theta_0)\right)$ in the formula of the asymptotic version of the KLM test, whereas, conditional on $\text{rk}(\theta_0)$, the bootstrapped CLR test is computed from the bootstrapped realizations of the AR and KLM tests.

S.5.2 Single-equation residual bootstrap for the KLM and CLR tests

Conditional on $\widetilde{\Pi}_z(\theta_0)$ and $\text{rk}(\theta_0)$, the KLM and CLR SE bootstrap test statistics are obtained in the same way as their EE bootstraps. Here, the bootstrap values of $\tilde{\lambda}_{\text{KLM}}(\theta_0)$ and $\widehat{\text{Var}}\left(\tilde{\lambda}_{\text{KLM}}(\theta_0)\right)$ are, respectively,

$$\begin{aligned} \acute{\lambda}^*_{\text{KLM}}(\theta_0) &= \widetilde{\Pi}_z(\theta_0)' \left[\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)\right]^{-1} \acute{\delta}^*_z(\theta_0), \text{ and} \\ \widehat{\text{Var}}\left(\acute{\lambda}^*(\theta_0)\right) &= \widetilde{\Pi}_z(\theta_0)' \left[\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)\right]^{-1} \widetilde{\Pi}_z(\theta_0). \end{aligned}$$

where $\acute{\Omega}^*_{\delta^*_z\delta^*_z}(\theta_0)$ and $\acute{\delta}^*_z(\theta_0)$ are obtained from the AR SE described in Subsection 5.2.

S.5.3 Multi-equation residual bootstraps

A simple procedure for obtaining bootstrap samples using the first- and second-stage residuals together is

$$\begin{cases} \acute{\mathbf{Y}}^*_g(\theta_0) = & \mathbf{w}_g\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}^*_g(\theta_0) \\ \hat{\mathbf{y}}^*_{2,g} = & \mathbf{w}_g\widehat{\Pi}_w + \hat{\mathbf{v}}^*_g, \end{cases}$$

where $\left\{\acute{\mathbf{e}}^*_g(\theta_0), \hat{\mathbf{v}}^*_g\right\}_{g=1}^G = \left\{\omega_g\acute{\mathbf{e}}_g(\theta_0), \omega_g\hat{\mathbf{v}}_g\right\}_{g=1}^G$, $\acute{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{w}_g\acute{\delta}_w(\theta_0)$ and $\hat{\mathbf{v}}_g = \mathbf{y}_{2,g} - \mathbf{w}_g\widehat{\Pi}_w$.

If the errors are homoskedastic and sampled at the individual level, Davidson and MacKinnon (2010) propose replacing $\widehat{\Pi}_w$ and $\hat{\mathbf{v}}_g$ by, respectively, $\breve{\Pi}_w(\theta_0)$ and $\breve{\mathbf{v}}_g(\theta_0)$, where $\breve{\Pi}_w(\theta_0)$ is the OLS estimator of Π_w obtained from the following auxiliary regression model

$$\mathbf{y}_2 = \mathbf{W}\Pi_w + \acute{\mathbf{e}}(\theta_0)\Gamma + \text{residuals},$$

and $\breve{\mathbf{v}}(\theta_0) = \mathbf{y}_2 - \mathbf{W}\breve{\Pi}_w(\theta_0)$. Davidson and MacKinnon show that this bootstrap DGP is a more efficient one because it incorporates information about the correlation between $\mathbf{e}_g$ and $\mathbf{v}_g$ when estimating the first-stage residuals.[5]

[5]If the residuals are homoskedastic, then $\breve{\Pi}_w(\theta_0)$ would be equivalent to the three-stage least squares (3SLS) or limited-information maximum likelihood estimator (under residuals join normality).

In the cluster case, however, we can use $\widetilde{\Pi}_w(\theta_0)$, the estimator of Π_w derived from (CS 4), and $\tilde{\delta}_w(\theta_0) = (0, \tilde{\gamma}(\theta_0))$, where $\tilde{\gamma}(\theta_0)$ is the continuous updating estimator derived from Equation (CS 4), in place of $\widehat{\Pi}_w$ and $\acute{\delta}_w(\theta_0)$. A new bootstrap residual method, therefore, uses $\tilde{\mathbf{e}}_g(\theta_0) = \mathbf{Y}_g(\theta_0) - \mathbf{x}_g\tilde{\gamma}(\theta_0)$ and $\tilde{\mathbf{v}}_g(\theta_0) = \mathbf{y}_{2,g} - \mathbf{w}_g\widetilde{\Pi}_w(\theta_0)$ in place of $\acute{\mathbf{e}}_g(\theta_0)$ and $\hat{\mathbf{v}}_g$, respectively, for the bootstrap DGP.

Definition 2 (Multi-equation residual (ME) bootstrap)**.** Let $\{\omega_g\}_{g=1}^G$ be a sequence of bootstrap weights satisfying $E[\omega_g] = 0$ and $\mathrm{Var}(\omega_g) = 1$. The bootstrap data generating process for $\{\mathbf{Y}_g(\theta_0), \mathbf{y}_{2,g}\}_{g=1}^G$ is:

1. Inefficient ME (ME-in):
$$\left\{\acute{\mathbf{Y}}_g^*(\theta_0), \hat{\mathbf{y}}_{2,g}^*\right\}_{g=1}^G = \left\{\mathbf{w}_g\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}_g^*(\theta_0), \mathbf{w}_g\widehat{\Pi}_w + \hat{\mathbf{v}}_g^*\right\}_{g=1}^G, \text{ where}$$
$$\left\{\acute{\mathbf{e}}_g^*(\theta_0), \hat{\mathbf{v}}_g^*\right\}_{g=1}^G = \left\{\omega_g\acute{\mathbf{e}}_g(\theta_0), \omega_g\hat{\mathbf{v}}_g\right\}_{g=1}^G;$$

2. Efficient ME (ME-eff):
$$\left\{\tilde{\mathbf{Y}}_g^*(\theta_0), \tilde{\mathbf{y}}_{2,g}^*(\theta_0)\right\}_{g=1}^G = \left\{\mathbf{x}_g\tilde{\gamma}(\theta_0) + \tilde{\mathbf{e}}_g^*(\theta_0), \mathbf{w}_g\widetilde{\Pi}_w(\theta_0) + \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G, \text{ where}$$
$$\left\{\tilde{\mathbf{e}}_g^*(\theta_0), \tilde{\mathbf{v}}_g^*(\theta_0)\right\}_{g=1}^G = \left\{\omega_g\tilde{\mathbf{e}}_g(\theta_0), \omega_g\tilde{\mathbf{v}}_g(\theta_0)\right\}_{g=1}^G.$$

Remark S.5.1. Since the AR test does not depend on the first-stage equation, the SE-in and ME-in bootstraps are equivalent. Similarly, the SE-eff and ME-eff bootstraps are equivalent.

Remark S.5.2. In contrast with the EE and SE bootstraps, $\widetilde{\Pi}_z(\theta_0)$ needs to be computed for every bootstrap sample when computing the KLM ME bootstrap test. The ME bootstrap version of CLR test, as explained in Davidson and MacKinnon (2008), requires a double bootstrap procedure: one bootstrap for the rank statistic and a second bootstrap conditional on the bootstrapped rank statistic. Therefore, we have not computed the ME version of this statistic.

S.5.4 Davidson and MacKinnon's bootstrap for linear IV models

The DM bootstrap method is a residual bootstrap based on the original AR and KLM tests, which are

$$\mathrm{AR}(\theta_0) = \frac{n-k_w}{k_z}\frac{\mathbf{Y}(\theta_0)'\,\mathrm{M}_{\mathbf{X}}\mathbf{Z}(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z})^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Y}(\theta_0)}{\mathbf{Y}(\theta_0)'\,\mathrm{M}_{\mathbf{W}}\mathbf{Y}(\theta_0)}, \quad \text{and} \tag{CS 8}$$

$$\mathrm{KLM}(\theta_0) = (n-k_w)\frac{\mathbf{Y}(\theta_0)'\,\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\acute{\Pi}_w(\theta_0)}\mathbf{Y}(\theta_0)}{\mathbf{Y}(\theta_0)'\,\mathrm{M}_{\mathbf{W}}\mathbf{Y}(\theta_0)}. \tag{CS 9}$$

Bootstrap realizations of $(\mathbf{y}_1, \mathbf{y}_2)$ are obtained from residuals sampled at the individual level.

When errors are homoskedastic, the AR and KLM tests defined on Equations (CS 8) and (CS 9) are distributed asymptotically as, respectively, $F_{(\infty,k_z)}$ and $\chi^2(p)$. These tests, however, are not pivotal when errors are heteroskedastic. Davidson and MacKinnon (2010) show that the limiting distributions of $n^{-\frac{1}{2}}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Y}(\theta_0)$ and $n^{-\frac{1}{2}}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Y}^*(\theta_0)$, where $\mathbf{Y}^*(\theta_0)$ is the bootstrap realization of $\mathbf{Y}(\theta_0)$, are equal. The same is true for the probability limits of $n^{-1}\mathbf{Y}(\theta_0)'\mathrm{M}_{\mathbf{W}}\mathbf{Y}(\theta_0)$ and $n^{-1}\mathbf{Y}^*(\theta_0)'\mathrm{M}_{\mathbf{W}}\mathbf{Y}^*(\theta_0)$. Therefore, the AR bootstrap test is valid under heterogeneity because it converges to the same limit distribution as the asymptotic AR test, irrespective of the instrument strength. This argument extends to the KLM bootstrap test, but only when the concentration parameter is high. Davidson and MacKinnon (2010) name their bootstrap method the *wild restricted efficient (WRE) residual bootstrap*.

We define the Davidson and MacKinnon (DM) bootstrap method as the WRE bootstrap but with residuals sampled at cluster level instead of individual level. The bootstrap sample is, therefore,

$$\left\{\acute{\mathbf{Y}}^*_g(\theta_0), \acute{\mathbf{y}}^*_{2,g}\right\}_{g=1}^{G} = \left\{\mathbf{w}_g\acute{\delta}_w(\theta_0) + \acute{\mathbf{e}}^*_g(\theta_0), \mathbf{w}_g\acute{\Pi}_w(\theta_0) + \acute{\mathbf{v}}^*_g(\theta_0)\right\}_{g=1}^{G},$$

where $\left\{\acute{\mathbf{e}}^*_g(\theta_0), \acute{\mathbf{v}}^*_g(\theta_0)\right\}_{g=1}^{G} = \left\{\omega_g\acute{\mathbf{e}}_g(\theta_0), \omega_g\acute{\mathbf{v}}_g(\theta_0)\right\}_{g=1}^{G}$, with $\acute{\mathbf{e}}_g(\theta_0) = \mathbf{y}_{1,g} - \mathbf{w}_g\acute{\delta}_w(\theta_0)$ and $\acute{\mathbf{v}}_g(\theta_0) = \mathbf{y}_{2,g} - \mathbf{w}_g\acute{\Pi}_w(\theta_0)$. We can use the same arguments in Davidson and MacKinnon (2010) to show that the DM bootstrap for the AR test with cluster residuals is consistent, assuming that the number of clusters is increasing and the number of observations within clusters is constant.

The DM bootstrap tests, however, do not perform satisfactorily in small samples with high degrees of heterogeneity. In order to illustrate this, we simulate the baseline model of 400 observations with 20 clusters of the same size, 5 instruments, normal errors, and a noncentrality parameter μ_{k_z} of 18. We change the skedastic function by removing the scaling factor in (12), which now becomes

$$\breve{f}\left(\mathbf{z}_{1,g},\kappa\right)=\left(\boldsymbol{\iota}_g+2\mathbf{z}_{1,g}\right)^{\kappa}. \tag{CS 10}$$

Table S.11 in Section S.8 shows that the DM bootstrap tests rejection rates are close to the nominal level of 5% under cluster random effect errors framework ($\kappa = 0$). In the presence of strong heteroskedastic residuals ($\kappa = 2$), the rejection rates of the DM bootstrap tests are above the rejection rates of asymptotic tests. In the same table, the proposed EE, SE-eff, and ME-eff bootstraps for the AR and KLM tests have rejection rates close to the nominal size.

S.5.5 Wald bootstrap tests

We investigate three methods for bootstrapping the Wald test. The first one, which is similar to the multi-equation residual bootstrap described at Subsection S.5.3, the bootstrap residuals are obtained from

$$\left\{\hat{\mathbf{u}}_g^*\left(\hat{\theta}_{\mathrm{IV}}\right),\hat{\mathbf{v}}_g^*\right\}_{g=1}^{G}=\left\{\omega_g\hat{\mathbf{u}}_g\left(\hat{\theta}_{\mathrm{IV}}\right),\omega_g\hat{\mathbf{v}}_g\right\}_{g=1}^{G},$$

where $\hat{\mathbf{u}}_g\left(\hat{\theta}_{\mathrm{IV}}\right)$ and $\hat{\mathbf{v}}_g$ are the IV and first-stage OLS residuals, respectively. Let $\hat{\theta}_{\mathrm{IV}}^*$ be the IV estimate obtained from the bootstrap sample. The Wald multi-equation instrumental variable (ME-IV) bootstrap test is

$$(\hat{\theta}_{\mathrm{IV}}^*-\hat{\theta}_{\mathrm{IV}})'\left(\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}}^*)\right)^{-1}(\hat{\theta}_{\mathrm{IV}}^*-\hat{\theta}_{\mathrm{IV}}),$$

where $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}}^*)$ is the variance estimate of $\hat{\theta}_{\mathrm{IV}}^*$ computed in the same way as $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}})$ in Equation (6).

We can also use the DGP of the ME efficient procedure for bootstrapping the Wald test. In this case the Wald bootstrap test is computed with $(\acute{\theta}_{\mathrm{IV}}^*-\theta_0)$ and $\widehat{\mathrm{Var}}(\acute{\theta}_{\mathrm{IV}}^*)$ in place of respectively $(\hat{\theta}_{\mathrm{IV}}^*-\hat{\theta}_{\mathrm{IV}})$ and $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{IV}}^*)$, where

$$\acute{\theta}_{\mathrm{IV}}^*=\left(\tilde{\mathbf{y}}_2^*\left(\theta_0\right)'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\tilde{\mathbf{y}}_2^*\left(\theta_0\right)\right)^{-1}\tilde{\mathbf{y}}_2^*\left(\theta_0\right)'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\acute{\mathbf{y}}_1^*\left(\theta_0\right),$$

with $\acute{\mathbf{y}}_1^*\left(\theta_0\right)=\tilde{\mathbf{y}}_{2,g}^*\left(\theta_0\right)\times\theta_0+\mathbf{x}\tilde{\gamma}\left(\theta_0\right)+\acute{\mathbf{e}}^*\left(\theta_0\right)$, and $\widehat{\mathrm{Var}}(\acute{\theta}_{\mathrm{IV}}^*)$ is a suitable variance estimate of $\acute{\theta}_{\mathrm{IV}}^*$. This is the IV bootstrap, which is based on samples generated under the null assumption, similar to the bootstraps used in the main manuscript. In our simulations, they have better size properties than the remaining Wald bootstrap methods presented in this subsection.

The third Wald bootstrap method is the classical *pairs bootstrap*. This method is a completely nonparametric one. In the pairs bootstrap, the bootstrap sample is generated as

$$\Pr\left(\left\{\mathbf{y}_{1,g}^{*}, \mathbf{y}_{2,g}^{*}, \mathbf{w}_{g}^{*}\right\} = \left\{\mathbf{y}_{1,j}, \mathbf{y}_{2,j}, \mathbf{w}_{j}\right\}\right) = \frac{1}{G}, \quad j = 1, \ldots, G.$$

The pairs bootstrap data generating process does not impose the null hypothesis on the bootstrap samples. Therefore, this Wald bootstrap test is centered at the IV estimate.

Table S.2 shows a summary of the bootstrap methods investigated in the study.

[Table 2 about here.]

S.6 Bias of the cluster IV estimator for θ

In matrix notation, the cluster residual model system (3) is

$$\left\{\begin{array}{rl} \mathbf{y}_1 = & \mathbf{y}_2\theta + \mathbf{X}\gamma + \mathbf{u} \\ \mathbf{y}_2 = & \mathbf{Z}\Pi_z + \mathbf{X}\Pi_x + \mathbf{V} \end{array}\right., \qquad \mathrm{E}\left[\begin{array}{cc} \mathbf{u}\mathbf{u}' & \mathbf{u}\underline{\mathbf{v}}' \\ \underline{\mathbf{v}}\mathbf{u}' & \underline{\mathbf{v}}\underline{\mathbf{v}}' \end{array}\middle| \mathbf{W}\right] = \left[\begin{array}{cc} \Sigma_{\mathbf{uu}} & \Sigma_{\mathbf{uv}} \\ \Sigma_{\mathbf{vu}} & \Sigma_{\mathbf{vv}} \end{array}\right],$$

where, $\underline{\mathbf{v}} = \mathrm{vec}\left(\mathbf{V}\right)$, $\Sigma_{\mathbf{uu}} = \mathrm{diag}[\{\Sigma_{\mathbf{u}_g\mathbf{u}_g}\}_{g=1}^{G}]$, $\Sigma_{\mathbf{uv}} = [\Sigma_{\mathbf{uv}_1}, \ldots, \Sigma_{\mathbf{uv}_p}]'$, with $\Sigma_{\mathbf{uv}_j} = \mathrm{diag}[\{\Sigma_{\mathbf{v}_{j,g}\mathbf{u}_g}\}_{g=1}^{G}]$ for $j = 1, \ldots, p$, and $\Sigma_{\mathbf{v}_j\mathbf{v}_m} = \mathrm{diag}\left[\{\Sigma_{\mathbf{v}_{j,g}\mathbf{v}_{m,g}}\}_{g=1}^{G}\right]$ for $j, m = 1, \ldots, p$. Let $n = \sum_{g=1}^{G} n_g$ and assume that $\mathrm{rank}\left(\Pi_z\right) = p$. The IV estimator $\hat{\theta}_{\mathrm{IV}}$ can be written as

$$\hat{\theta}_{\mathrm{IV}} - \theta = \left(\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{y}_2\right)^{-1}\mathbf{y}_2'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{u} = \left(\mathrm{I} + \mathbf{Q}^{-1}\Delta\right)^{-1}\mathbf{Q}^{-1}\mathbf{C}, \tag{CS 11}$$

where $\mathbf{Q} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z$, $\Delta = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V} + \mathbf{V}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z + \mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{V}$, and $\mathbf{C} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{u} + \mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{u}$. We have $\mathbf{Q}^{-1} = n^{-1} \times O_p\left(1\right)$ and $\Delta = \sqrt{n} \times O_p\left(1\right)$, which implies that, as $n \longrightarrow +\infty$, $\left(\mathrm{I} + \mathbf{Q}^{-1}\Delta\right)^{-1} = O_p\left(1\right)$. By Taylor expansion around $\Pi_z = 0$, we derive $\left(\mathrm{I} + \mathbf{Q}^{-1}\Delta\right)^{-1} \approx \mathrm{I} - \mathbf{Q}^{-1}\Delta$. Therefore, Equation (CS 11) can be simplified to

$$\hat{\theta}_{\mathrm{IV}} - \theta = \mathbf{Q}^{-1}\left\{\mathbf{C}_{\mathbf{Vu}} - \left(\Delta_{\mathbf{V}} + \Delta_{\mathbf{V}}'\right)\mathbf{Q}^{-1}\mathbf{C}_{\mathbf{u}}\right\} + \mathbf{H},$$

where $\mathbf{C}_{\mathbf{Vu}} = \mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\mathbf{u}$, $\Delta_{\mathbf{V}} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{V}$, $\mathbf{C}_{\mathbf{u}} = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{u}$, and $\mathbf{H}$ has terms related to odd moments of the joint distribution of $\left(\mathbf{u}', \underline{\mathbf{v}}'\right)'$ and terms which are of small order.

Assuming that the first and third moments of the joint distribution are equal to zero,[6] the

[6]If $\left(\mathbf{u}, \underline{\mathbf{v}}\right)$ follows a multivariate distribution, we can invoke Isselis's or Wick's theorem, which say that the expected value of odd moments of a centered multivariate normal distribution are 0.

bias of the IV estimator is, approximately,

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right] \approx \mathrm{E}\left\{\mathbf{Q^{-1}}\left[\mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}-\mathbf{\Delta}_{\mathbf{V}}\mathbf{Q}^{-1}\mathbf{C}_{\mathbf{u}}\right]\right\}, \tag{CS 12}$$

where $\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}} \equiv \mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}-\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z}$. The first element of $\mathbf{V}'\,\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}$ is $\mathbf{v}_1'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}$, whose expectation is $\mathrm{E}[\mathbf{v}_1'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}] = \mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_1})$. So, $\mathrm{E}[\mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}] = [\mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_1}),$ $\ldots, \mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_p})]'$. Partition $\mathbf{Q}^{-1}$ as $\mathbf{Q}^{-1}=[\mathbf{Q}^{\cdot 1},\ldots,\mathbf{Q}^{\cdot p}]$. Therefore, we obtain

$$\mathrm{E}\left[\mathbf{Q^{-1}}\mathbf{V}'\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\mathbf{u}\right] = \sum_{j=1}^{p}\mathrm{trace}\left(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_j}\right)\mathbf{Q}^{\cdot j}. \tag{CS 13}$$

To study $\mathrm{E}[\mathbf{Q^{-1}}\mathbf{\Delta}_{\mathbf{V}}\mathbf{Q}^{-1}\mathbf{C}_{\mathbf{u}}]$, we rewrite $\mathbf{\Delta}_{\mathbf{V}}\mathbf{Q}^{-1}\mathbf{C}_{\mathbf{u}}$ as

$$\mathrm{vec}(\mathbf{C}_{\mathbf{u}}'\mathbf{Q}^{-1}\mathbf{\Delta}_{\mathbf{V}}') = \Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}(\mathbf{V}\otimes\mathbf{u}')(\mathrm{I}_p\otimes\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z)\,\mathrm{vec}(\mathbf{Q}^{-1}).$$

Since $\mathrm{E}(\mathbf{V}\otimes\mathbf{u}')=[\Sigma_{\mathbf{v}_1\mathbf{u}},\ldots,\Sigma_{\mathbf{v}_p\mathbf{u}}]$, we have $\mathrm{E}[\mathbf{C}_{\mathbf{u}}'\mathbf{Q}^{-1}\mathbf{\Delta}_{\mathbf{V}}'] = \sum_{j=1}^{p}\Pi_z'\,\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\,\Sigma_{\mathbf{v}_j\mathbf{u}}\,\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z\mathbf{Q}^{\cdot j}$, and, consequently,

$$\mathrm{E}\left[\mathbf{Q}^{-1}\mathbf{\Delta}_{\mathbf{V}}\mathbf{Q}^{-1}\mathbf{C}_{\mathbf{u}}\right] = \sum_{j=1}^{p}\mathbf{Q}^{-1}\Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{v}_j\mathbf{u}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z\mathbf{Q}^{\cdot j}. \tag{CS 14}$$

Substituting Equations (CS 13) and (CS 14) into (CS 12) we obtain

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right] \approx \sum_{j=1}^{p}\mathrm{trace}\left(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z^{\perp}}\Sigma_{\mathbf{uv}_j}\right)\mathrm{I}_p\mathbf{Q}^{\cdot j} - \sum_{j=1}^{p}\mathbf{Q}^{-1}\Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{v}_j\mathbf{u}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z\mathbf{Q}^{\cdot j}. \tag{CS 15}$$

When $p=1$, Equation (CS 15) simplifies to $\mathrm{trace}(\mathrm{P}_{\mathrm{M}_{\mathbf{X}}\mathbf{Z}}\Sigma_{\mathbf{uv}})\mathbf{Q^{-1}} - 2(\Pi_z'\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{uv}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}\Pi_z)\mathbf{Q}^{-2}$, which is the same as the bias derived by Bun and de Haan (2010). Assuming that $(\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\mathbf{Z}) = n\mathrm{I}_{k_z}$ and $\Pi_z = \|\Pi_z\|\,\Pi_0$, we find:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right] \approx \mu^{-2}\frac{\mathrm{trace}\,(\mathrm{S}_{12})}{\mathrm{trace}\,(\mathrm{S}_{22})}\left\{1-2\frac{(\Pi_0'\mathrm{S}_{12}\Pi_0)}{\mathrm{trace}\,(\mathrm{S}_{12})}\right\}$$

where $\mathrm{S}_{12} = n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{uv}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, $\mathrm{S}_{22} = n^{-1}\mathbf{Z}'\mathrm{M}_{\mathbf{X}}\Sigma_{\mathbf{vv}}\mathrm{M}_{\mathbf{X}}\mathbf{Z}$, and $\mu^2 = [\mathrm{trace}\,(\mathrm{S}_{22})]^{-1}\,n\,\|\Pi_z\|^2$, which plays similar role as the concentration parameter (see Olea and Pflueger's Theorem 1). If the residuals are homoskedastic, then $\Sigma_{\mathbf{uv}_j} = \tau_j\mathrm{I}_n$, and Equation (CS 15) is simplified to $(k_z-p-1)\,\mathbf{Q}^{-1}\boldsymbol{\tau}$, where $\tau'=[\tau_1,\ldots,\tau_p]$.

The errors of the model are generated as $\mathbf{u}_g = \iota_{n_g}\upsilon_g+\varepsilon_g$ and $\mathrm{vec}(\mathbf{v}_g) = \left[\boldsymbol{\nu}_g\otimes\iota_{n_g}\right]+\boldsymbol{\epsilon}_g$ for

$g = 1,\ldots,G$.The random variable υ_g is scalar, and $\boldsymbol{\nu}_g$ and $\boldsymbol{\epsilon}_g$ are $p\times 1$ and $(n_g p)\times 1$ vectors with $(\upsilon_g,\boldsymbol{\nu}_g)' \sim \sqrt{\phi_g}N(0,\ [1,\ \boldsymbol{\rho}'\ :\ \boldsymbol{\rho}],\ \mathrm{I}_p)$, $(\varepsilon_g,\boldsymbol{\epsilon}_g)' \sim \sqrt{\varphi_g}N(0,[1,\boldsymbol{\varrho}'\ :\ \boldsymbol{\varrho},\mathrm{I}_p]\otimes\mathrm{I}_{n_g})$, and ϕ_g is a scalar such that $1\geq\phi_g\geq 0$. The $p\times 1$ correlation vectors $\boldsymbol{\rho}' = [\rho_1,\ldots,\rho_p]$ and $\boldsymbol{\varrho}' = [\varrho_1,\ldots,\varrho_p]$ capture the intra-cluster and the idiosyncratic on the endogeneity degree. Under these assumptions, the joint distribution of $\left(\mathbf{u}_g',\mathrm{vec}(\mathbf{v}_g)'\right)'$ is:

$$\left(\begin{array}{c}\mathbf{u}_g\\ \mathrm{vec}(\mathbf{v}_g)\end{array}\right)\sim N\left(0,\left[\begin{array}{cc}\mathsf{W}_g+\overline{\mathsf{W}}_g & \left(\boldsymbol{\rho}'\otimes\mathrm{I}_{n_g}\right)\mathsf{W}_g+\left(\boldsymbol{\varrho}'\otimes\mathrm{I}_{n_g}\right)\overline{\mathsf{W}}_g\\ \cdot & \mathrm{I}_p\otimes\left(\mathsf{W}_g+\overline{\mathsf{W}}_g\right)\end{array}\right]\right)$$

where $\mathsf{W}_g = \phi_g\ \iota_{n_g}\iota_{n_g}'$, $\overline{\mathsf{W}}_g = (1-\phi_g)\ \mathrm{I}_{n_g}$. We interpret ϕ_g and $(1-\phi_g)$ as the weights due to the cluster and idiosyncratic effects in the correlation. We can rewrite $\Sigma_{\mathbf{uv}_j}$ and $\mathbf{Z}'\mathrm{M}_\mathbf{X}\Sigma_{\mathbf{uv}_j}\mathrm{M}_\mathbf{X}\mathbf{Z}$ as $\Sigma_{\mathbf{uv}_j} = \mathrm{diag}\{\ \rho_j\mathsf{W}_g + \varrho_j\overline{\mathsf{W}}_g\}_{g=1}^G$ and $\mathbf{Z}'\mathrm{M}_\mathbf{X}\Sigma_{\mathbf{uv}_j}\mathrm{M}_\mathbf{X}\mathbf{Z} = \mathbf{Z}'\mathrm{M}_\mathbf{X}\left(\rho_j\mathsf{W}+\ \varrho_j\overline{\mathsf{W}}\right)\mathrm{M}_\mathbf{X}\mathbf{Z}$ for $j = 1,\ldots,p$, where $\mathsf{W} = \mathrm{diag}\{\mathsf{W}_g\}_{g=1}^G$, and $\overline{\mathsf{W}} = \mathrm{diag}\left\{\overline{\mathsf{W}}_g\right\}_{g=1}^G$. Then, after further simplifications, the bias of IV estimator turns out to be

$$\begin{aligned}\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right]\approx&\left\{\mathrm{trace}\left[\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}\Pi_z^\perp}\mathsf{W}\right]\mathrm{I}_p-\mathbf{Q}^{-1}\Pi_z'\left(\mathbf{Z}'\mathrm{M}_\mathbf{X}\mathsf{W}\mathrm{M}_\mathbf{X}\mathbf{Z}\right)\Pi_z\right\}\mathbf{Q^{-1}}\boldsymbol{\rho}\\&+\left\{\mathrm{trace}\left[\mathrm{P}_{\mathrm{M}_\mathbf{X}\mathbf{Z}\Pi_z^\perp}\overline{\mathsf{W}}\right]\mathrm{I}_p-\mathbf{Q}^{-1}\Pi_z'\left(\mathbf{Z}'\mathrm{M}_\mathbf{X}\overline{\mathsf{W}}\mathrm{M}_\mathbf{X}\mathbf{Z}\right)\Pi_z\right\}\mathbf{Q^{-1}}\boldsymbol{\varrho}\end{aligned}$$

The first term captures the bias of the IV estimator due to the cluster effect while the second term is a function of the within cluster correlations. Let us define $\mathbf{Z}$ and $\mathrm{M}_\mathbf{X}\mathbf{Z}$ as $\mathbf{Z} = \left[\mathbf{d}_1'\iota_{n_1}'+\vartheta_1',\ldots,\mathbf{d}_G'\iota_{n_G}'+\vartheta_G'\right]'$ and $\mathrm{M}_\mathbf{X}\mathbf{Z} = \left[\mathbf{z_1^\perp}',\ldots,\mathbf{z_G^\perp}'\right] = [\left(\mathbf{d}_1-\overline{\mathbf{d}}\right)'\iota_{n_1}'+\left(\vartheta_1-\iota_{n_1}\overline{\vartheta}\right)'$, ..., $\left(\mathbf{d}_G-\overline{\mathbf{d}}\right)'\iota_{n_G}'+\left(\vartheta_G-\iota_{n_G}\overline{\vartheta}\right)']'$ where $\overline{\mathbf{d}} = \left(n^{-1}\sum_{g=1}^G n_g\mathbf{d}_g\right)$and $\overline{\vartheta} = \left(n^{-1}\sum_{g=1}^G\iota_{n_g}'\vartheta_g\right)$. We interpret $\mathbf{d}_g$ and the part of instruments which is common to all observations in cluster g, while that ϑ_g captures the part which is idiosyncratic for each observation. Define $\overline{\vartheta}_g = n_g^{-1}\iota_{n_g}'\vartheta_g$. If we further impose in the data generate process that $\overline{\vartheta}_g = \overline{\vartheta} = 0$, $\phi_g = \phi$, $n_g = \bar{n}$ for all g, then we have $\mathbf{Z}'\mathrm{M}_\mathbf{X}\mathbf{Z} =\bar{n}\sum_{g=1}^G[\left(\mathbf{d}_g-\overline{\mathbf{d}}\right)'\left(\mathbf{d}_g-\overline{\mathbf{d}}\right)+\vartheta_g'\vartheta_g]$, $\mathbf{Z}'\mathrm{M}_\mathbf{X}\overline{\mathsf{W}}\mathrm{M}_\mathbf{X}\mathbf{Z}=(1-\phi)\,\mathbf{Z}'\mathrm{M}_\mathbf{X}\mathbf{Z}$, and $\mathbf{Z}'\mathrm{M}_\mathbf{X}\mathsf{W}\mathrm{M}_\mathbf{X}\mathbf{Z}=\phi\bar{n}\left(\mathbf{Z}'\mathrm{M}_\mathbf{X}\mathbf{Z}-\sum_{g=1}^G\vartheta_g'\vartheta_g\right)$. When generating the data, we rescale the values of $\mathbf{d}_g$ such that $\sum_{g=1}^G n_g\left(\mathbf{d}_g-\overline{\mathbf{d}}\right)'\left(\mathbf{d}_g-\overline{\mathbf{d}}\right) = (1-\lambda)\,n\mathrm{I}_{k_z}$ and $(\sum_{g=1}^G\vartheta_g'\vartheta_g) = \lambda n\mathrm{I}_{k_z}$, so $\mathbf{Z}'\mathrm{M}_\mathbf{X}\mathbf{Z}$ $= n\mathrm{I}_{k_z}$ for $0\leq\lambda\leq 1$. Then, the bias of IV estimator becomes

$$\begin{aligned}\mathrm{E}\left[\hat{\theta}_{\mathrm{IV}}-\theta\right]\approx&\quad\phi\bar{n}\,(1-\lambda)\,(k_z-p-1)\,\mathbf{Q^{-1}}\boldsymbol{\rho}+(1-\phi)\,(k_z-p-1)\,\mathbf{Q^{-1}}\boldsymbol{\varrho}\\\approx&\quad(k_z-p-1)\,\mathbf{Q^{-1}}\left[\phi\bar{n}\,(1-\lambda)\,\boldsymbol{\rho}+(1-\phi)\,\boldsymbol{\varrho}\right]\end{aligned}$$

The last equation is the same as Equation (15).

S.7 The first-stage F- and effective F-tests

If there is one endogenous variable, the first-stage F-statistic for the null hypothesis $H_0 : \Pi_z = 0$ is defined as

$$\hat{\mathrm{F}} \equiv \frac{\hat{\Pi}_z' \left(\widehat{\mathrm{Var}}(\hat{\Pi}_z) \right)^{-1} \hat{\Pi}_z}{k_z}, \tag{CS 16}$$

where $\widehat{\mathrm{Var}}(\hat{\Pi}_z)$ is the cluster-robust variance matrix. Testing instrument weakness in the cluster IV model can be performed by the *effective F-test*, see Olea and Pflueger (2013). It is defined as

$$\hat{\mathrm{F}}_{eff} \equiv \frac{\mathbf{y}_2' \mathbf{Z}^{\perp} \mathbf{Z}^{\perp\prime} \mathbf{y}_2}{n \,\mathrm{tr}\left(\hat{\mathrm{S}}_{22}\right)}, \tag{CS 17}$$

with *effective degrees of freedom*

$$\hat{k}_{eff} \equiv \frac{\left[\mathrm{tr}\left(\hat{\mathrm{S}}_{22}\right)\right]^2 \left(1 + 2\hat{f}\right)}{\mathrm{tr}\left(\hat{\mathrm{S}}_{22}' \hat{\mathrm{S}}_{22}\right) + 2\hat{f}\,\mathrm{tr}\left(\hat{\mathrm{S}}_{22}\right) \max\mathrm{eval}\left(\hat{\mathrm{S}}_{22}\right)},$$

where $\mathbf{Z}^{\perp} = (\mathbf{Z}\mathrm{M}_{\mathbf{X}}\mathbf{Z})^{-\frac{1}{2}} \mathrm{M}_{\mathbf{X}}\mathbf{Z}$, $\hat{\mathrm{S}}_{22}$ is the lower $k_z \times k_z$ submatrix of $\hat{\mathrm{S}}$, defined as

$$\hat{\mathrm{S}} = A \hat{\Xi} A', A = \begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix} \otimes \mathrm{I}_{k_z},$$

where $\hat{\Xi}$ is the estimator of Ξ, the asymptotic variance of $n^{-\frac{1}{2}} \mathbf{Z}^{\perp\prime} (\mathbf{e}, \mathbf{V})'$, $\mathrm{tr}(\cdot)$ and $\max\mathrm{eval}(\cdot)$ are the trace and the maximum eigenvalue operators, respectively, and $\hat{f}$ is an estimator of

$$f = \frac{\mathrm{B}_{\mathrm{IV}}(\Xi, \Omega)}{\tau}$$

where τ represents a tolerance parameter. The scalar $\mathrm{B}_{\mathrm{IV}}(\Xi, \Omega)$ is defined as

$$\mathrm{B}_{\mathrm{IV}}(\Xi, \Omega) \equiv \sup_{\beta \in \mathbb{R}, \Pi_0 \in S^{k-1}} \frac{|\mathrm{b}(\Xi, \beta, \Pi_0)|}{\mathrm{BM}(\Xi, \beta)}, \qquad \mathrm{B}_{\mathrm{IV}}(\Xi, \Omega) \le 1,$$

where S^{k-1} is a $k_z - 1$ dimension unit sphere.

Let partition matrix S as S= $[S_{11}, S_{12} : S_{21}, S_{22}]$. Then, $b(\Xi, \beta, \Pi_0)$ is defined as

$$b(\Xi, \beta, \Pi_0) = \left(\frac{\text{trace}(S_{12})}{\text{trace}(S_{22})} \left\{ 1 - 2\frac{(\Pi_0' S_{12} \Pi_0)}{\text{trace}(S_{12})} \right\} \right),$$

and $BM(\Xi, \beta) = \sqrt{\frac{S_{11}}{S_{22}}}$. The benchmark BM (Ξ, β) is interpreted as the worst case IV estimator bias. The Nagar bias is $\mu^2 \times b(\Xi, \beta, \Pi_0)$, where $\mu^2 = [\text{trace}(S_{22})]^{-1} n \|\Pi_z\|^2$. The null hypothesis of instrument's weakness is $H_0 : \mu^2 \times B_{IV}(\Xi, \Omega) > \tau$, which indicates that the relative bias of the IV estimator is greater than a certain tolerance level.

The estimation of $\hat{k}_{eff}$ requires a numerical algorithm. We translated the `Stata` code of Pflueger and Wang (2015) into `Matlab`. The simplified version for estimating $\hat{k}_{eff}$ sets $B_{IV}(\Xi, \Omega) = 1$. Therefore, for a tolerance of 10% we set f equal to 10. In our simulation results, we report both versions of the effective F tests. The rejection rates of the simplified effective F-test are only slightly smaller than the generalized one in the majority of cases.

When $p = 1$, the F-first stage on Equation (CS 16) can be asymptotically approximated by $\mu_{k_z} + F(k_z, +\infty)$ under the alternative hypothesis $H_1 \neq 0$, where $\mu_{k_z} = \Pi_z' \left[k_z \, \text{Var}_\infty(\hat{\Pi}_z) \right]^{-1} \Pi_z$ is the noncentrality parameter, and $F(k_z, +\infty)$ represents the asymptotic F-distribution.

The parameter μ_{k_z} is closely related to the measure μ^2. In our baseline simulation ($\kappa = 0$, $\eta = 0$), $n^{-1}E[\mathbf{Z}' M_\mathbf{X} \mathbf{V}\mathbf{V}' M_\mathbf{X} \mathbf{Z}]$ is, under our assumptions about the distribution of the errors, $n^{-1}\xi(\phi, \bar{n}, \lambda)\ \mathbf{Z}' M_\mathbf{X} \mathbf{Z}$, where $\xi(\phi, \bar{n}, \lambda) = \phi\bar{n}(1-\lambda) + (1-\phi)$. Then, the $\hat{F}_{eff}$ can be approximated to

$$\frac{\Pi_z'(\mathbf{Z}' M_\mathbf{X} \mathbf{Z})\Pi_z}{\xi(\phi, \bar{n}, \lambda)\, \text{trace}(n^{-1}\mathbf{Z}' M_\mathbf{X} \mathbf{Z})} + \frac{\mathbf{V}' M_\mathbf{X} \mathbf{Z}(\mathbf{Z}' M_\mathbf{X} \mathbf{Z})^{-1}(\mathbf{Z}' M_\mathbf{X} \mathbf{V})}{\xi(\phi, \bar{n}, \lambda)\, \text{trace}(n^{-1}\mathbf{Z}' M_\mathbf{X} \mathbf{Z})}.$$

Using the fact that $\text{Var}(\hat{\Pi}_z) = \xi(\phi, \bar{n}, \lambda)(\mathbf{Z}' M_\mathbf{X} \mathbf{Z})^{-1}$ and $n^{-1}\mathbf{Z}' M_\mathbf{X} \mathbf{Z} = I_{k_z}$, the later term becomes $\mu_{k_z} + {k_z}^{-1}\mathbf{V}' M_\mathbf{X} \mathbf{Z} \left[\text{Var}(\hat{\Pi}_z) \right]^{-1} (\mathbf{Z}' M_\mathbf{X} \mathbf{V})$. The second term of the above expression is asymptotically distributed as $F(k_z, +\infty)$.

The effective degrees of freedom is defined as

$$k_{eff} \equiv \frac{\left[\text{trace}\left(\frac{E[\mathbf{Z}' M_\mathbf{X} \mathbf{V}\mathbf{V}' M_\mathbf{X} \mathbf{Z}]}{n} \right) \right]^2 (1+2x)}{\text{trace}\left(\frac{E[\mathbf{Z}' M_\mathbf{X} \mathbf{V}\mathbf{V}' M_\mathbf{X} \mathbf{Z}]}{n} \frac{E[\mathbf{Z}' M_\mathbf{X} \mathbf{V}\mathbf{V}' M_\mathbf{X} \mathbf{Z}]}{n} \right) + 2\,\text{trace}\left(\frac{E[\mathbf{Z}' M_\mathbf{X} \mathbf{V}\mathbf{V}' M_\mathbf{X} \mathbf{Z}]}{n} \right) \max \text{eval}\left(\frac{E[\mathbf{Z}' M_\mathbf{X} \mathbf{V}\mathbf{V}' M_\mathbf{X} \mathbf{Z}]}{n} \right) x}$$

In our case, the numerator is simplified to $[\text{trace}(\xi(\phi, \bar{n}, \lambda)\, nI_{k_z})]^2 (1+2x)$, while the denominator becomes $\text{trace}\left([\xi(\phi, \bar{n}, \lambda)]^2 n^2 I_{k_z} \right) + 2\,\text{trace}(\xi(\phi, \bar{n}, \lambda)\, nI_{k_z}) \max \text{eval}(\xi(\phi, \bar{n}, \lambda)\, nI_{k_z})\, x$. Therefore, k_{eff} becomes

$$k_{eff} = \frac{\left(\xi\left(\phi, \bar{n}, \lambda\right) k_z\right)^2 (1+2x)}{\left(\xi\left(\phi, \bar{n}, \lambda\right)\right)^2 k_z + 2\left(\xi\left(\phi, \bar{n}, \lambda\right)\right)^2 k_z x} = \frac{k_z\,(1+2x)}{(1+2x)} = k_z.$$

S.8 Extra Simulation Results

We provide simulation results for the KLM and CLR asymptotic and bootstrap tests discussed in the main manuscript using the same simulation designs. We also report results for the DM bootstrap tests, and for the Wald ME-IV and pairs bootstraps. These bootstrap procedures are respectively described in Subsections S.5.4 and S.5.5. The rejection rates are computed by testing the true null assumption $H_0 : \theta = 0$ at 5% significance level. We also include the rejection rates of the first-stage F test as well as the conservative version of the effective F-test.

Tables S.3 and S.4 have results according to different distributions of the errors when instruments are strong ($\mu_{k_z} = 18$) and weak ($\mu_{k_z} = 0.1$), respectively. Tables S.5 and S.6 show results when the number of instruments increase, while Tables S.7 and S.8 contain the rejection rates when the number of clusters increases but the number of instruments are kept constant ($k_z = 5$). In all those tables, we simulate DGPs with a high degree of endogeneity ($\rho = 0.95$) and 20 observations per cluster. We note that:

1. When instruments are strong, the Wald multi-equation efficient (ME-eff) bootstrap, which is the bootstrap method with simulates the bootstrap samples imposing the null assumption, has superior size performance when compared with the other Wald bootstraps. The AR, KLM and CLR bootstraps have rejection rates close to the nominal values in almost all cases. The exception is the estimation equations bootstrap with multinomial weights and a high number of instruments ($k_z > 5$).

2. When instruments are extremely weak ($\mu_{k_z} = 0.1$), only the AR bootstrap tests rejection rates remain close to the nominal level for all levels of error heterogeneity (κ). Under weak instruments, the Wald bootstrap procedures can be as size distorted as the asymptotic Wald test.

Tables S.9 and S.10 highlight the role of endogeneity for different levels of instrument strength. In the strong instrument scenario (Table S.9), rejections rates are not sensitive to variations of ρ. Only when instruments are weak (Table S.10) do the Wald rejections rates increase together with the degree of endogeneity ρ.

Finally, Table S.11 shows simulation evidence of the size distortion of the DM bootstrap tests under severe heteroskedasticity. In the same table, we observe that the remaining AR and KLM bootstrap tests still have rejection rates close to the nominal level.

Figures S.1 to S.3 show the AR bootstrap test power curves for different degrees of error heteroskedasticity ($\kappa = 0$ to $\kappa = 2$) at $\mu_{k_z} = 18$. Finally, Figures S.4 to S.6 show the power curves for the bootstrapped Wald, KLM, and CLR tests when $\kappa = 1$ and $\mu_{k_z} = 18$.

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2012. The Colonial Origins of Comparative Development: An Empirical Investigation: Reply. *American Economic Review* 102(6): 3077–110. doi:10.1257/aer.102.6.3077

Andrews, Donald W.K. and Patrik Guggenberger. 2017. Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models. *Econometric Theory* 33(5): 1046–80. doi:10.1017/S0266466616000347

Bun, Maurice and Monique de Haan. 2010. Weak Instruments and the First Stage F-Statistic in IV Models With a Nonscalar Error Covariance Structure. UvA Econometrics Discussion Paper 2010/02.

Chernozhukov, Victor and Christian Hansen. 2008. The Reduced Form: A Simple Approach to Inference with Weak Instruments. *Economics Letters* 100(1): 68–71. doi:10.1016/j.econlet.2007.11.012

Davidson, Russell and James G. MacKinnon. 2008. Bootstrap Inference in a Linear Equation Estimated by Instrumental Variables. *Econometrics Journal* 11(3): 443–77. doi:10.1111/j.1368-423X.2008.00247.x

———. 2010. Wild Bootstrap Tests for IV Regression. *Journal of Business and Economic Statistics* 28(1): 128–44. doi:10.1198/jbes.2009.07221

Finlay, Keith and Leandro M. Magnusson. 2009. Implementing Weak Instrument Robust Tests for a General Class of Instrumental Variables Models. *Stata Journal* 9(3): 398–421.

Goldberger, Arthur and Ingram Olkin. 1971. A Minimum-Distance Interpretation of Limited-Information Estimation. *Econometrica* 39(3): 635–39.

Kleibergen, Frank and Richard Paap. 2006. Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics* 133(1): 97–126. doi:10.1016/j.jeconom.2005.02.011

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy* 112(4): 725–53. doi:10.1086/421174

Nagar, A. L. 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* 27(4): 575–95.

Olea, José Luis Montiel and Carolin Pflueger. 2013. A Robust Test for Weak Instruments. *Journal of Business and Economic Statistics* 31(3): 358–69. doi:10.1080/00401706.2013.806694

Pflueger, Carolin and Su Wang. 2015. A Robust Test for Weak Instruments in Stata. *Stata Journal* 15(1): 216–25.

Robin, Jean-Marc and Richard J. Smith. 2000. Tests of Rank. *Econometric Theory* 16(2): 151–75.

Table S.1: Application summary, 95% projection-based confidence intervals, First-stage robust F tests, and p-values ($\times 100$) for $H_0 : \theta_1 = \theta_2 = 0$, Miguel et al. (2004).

Source of dependent variable	95% projection-based confidence interval: Wald asym.	AR asym.	AR wild boot.	Ratio of CIs: AR /Wald	AR boot. /Wald	First Stage F-test
		θ_1				
PRIO/Uppsala	[-7.72 , 1.42]	[-13.10, 0.71]	$(-\infty, +\infty)$	1.51	$+\infty$	5.12
Doyle and Sambanis	[-4.25 , 1.00]	[-8.86 , 0.28]	$(-\infty$, 0.38]	1.74	$+\infty$	5.24
Fearon and Laitin	[-2.75 , 1.08]	[-3.59 , 1.44]	[-3.95 , 1.88]	1.31	1.52	5.18
		θ_2				
PRIO/Uppsala	[-5.47 , 1.79]	[-10.32, 29.90]	$(-\infty , +\infty)$	5.54	$+\infty$	3.43
Doyle and Sambanis	[-2.65 , 0.72]	[-3.25 , 9.96]	[-3.53 , $+\infty$)	3.92	$+\infty$	4.28
Fearon and Laitin	[-1.63 , -0.07]	[-4.43 , 2.11]	[-5.50 , 3.25]	4.18	5.60	4.20
$H_0 : \theta_1 = \theta_2 = 0$		p-values				
PRIO/Uppsala	0.179	0.101	0.114			
Doyle and Sambanis	0.261	0.122	0.140			
Fearon and Laitin	0.018	0.005	0.008			

Notes: The wild bootstraps use Rademacher resampling weights. The boxes indicate if the test is not rejected at the 1% significance level, the 5% level, the 10% level, and the 20% level. Mortality rates have been capped at 250 deaths per year per 1000 population.

Table S.2: Bootstrap methods for cluster IV

Method	Tests	Weights	Estimator for $\delta_w(\theta_0)$	Fixed $\widetilde{\Pi}_z(\theta_0)$?	H_0 imposed
Estimating Equations (EE)	AR, KLM, CLR	M, Γ, R	$\tilde{\delta}_w(\theta_0)$	Yes	Yes
Residuals Single Equation					
Inefficient (SE-in)	AR, KLM, CLR	Γ, R	$\acute{\delta}_w(\theta_0)$	Yes	Yes
Efficient (SE-eff)	AR, KLM, CLR	Γ, R	$\tilde{\delta}_w(\theta_0)$	Yes	Yes
Residuals Multiple Equation					
IV (ME-IV)	Wald	Γ, R		No	No
Inefficient (ME-in)	KLM	Γ, R	$\acute{\delta}_w(\theta_0)$	No	Yes
Efficient (ME-eff)	KLM, Wald	Γ, R	$\tilde{\delta}_w(\theta_0)$	No	Yes
Davidson-MacKinnon (DM)	AR, KLM	Γ, R	$\acute{\delta}_w(\theta_0)$	No	Yes
Pairs	Wald	M			No

Notes: The weights M, Γ, and R correspond to the multinomial, gamma, and Rademacher weights, respectively. The Wald ME-IV, Wald pairs, KLM ME-in, KLM ME-eff, and DM bootstraps are presented in the Supplementary material.

Table S.3: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 18$, endogeneity degree $\rho = 0.95$, varying error distributions

			$\mathcal{N}$ errors			χ^2 errors			t errors		
Stat.	Method	$\kappa \rightarrow$	0	1	2	0	1	2	0	1	2
Wald	Asymp.		33.70	41.88	45.73	34.75	41.71	50.10	31.12	41.02	46.54
	ME-IV	Γ	27.46	34.65	38.03	27.08	33.98	40.86	25.59	34.11	38.74
		R	30.01	37.60	41.44	30.24	37.72	45.20	28.29	37.37	42.56
	ME-eff	Γ	0.52	0.77	1.19	0.08	0.51	0.60	0.72	1.12	1.64
		R	1.42	2.06	2.40	0.53	1.24	1.46	2.06	2.27	2.97
	Pairs		10.58	18.00	22.26	7.13	17.27	24.62	11.21	19.37	24.57
AR	Asymp.		17.08	16.99	16.95	13.33	13.75	13.41	15.56	15.89	15.22
	EE	M	4.42	4.30	4.06	2.93	3.10	2.65	3.51	3.10	2.90
		Γ	6.30	6.23	6.17	4.75	5.24	4.68	5.51	5.54	5.46
		R	4.46	4.48	4.43	3.52	4.34	3.75	4.02	4.16	4.06
	SE-in	Γ	6.89	6.90	6.82	5.00	5.48	5.06	6.57	6.60	6.29
		R	5.39	5.28	5.24	3.87	4.44	3.89	4.99	5.20	5.04
	SE-eff	Γ	6.58	6.92	6.77	5.38	5.72	5.34	6.63	6.45	6.42
		R	5.07	5.00	5.00	3.98	4.41	4.08	4.67	4.72	4.63
	DM	Γ	2.69	1.76	0.54	2.25	1.40	0.66	2.26	1.54	0.73
		R	5.59	5.63	5.71	5.05	5.30	5.43	5.45	5.71	5.66
KLM	Asymp.		12.13	12.12	11.81	9.66	10.56	9.74	11.32	11.04	10.83
	EE	M	2.38	1.89	1.68	2.05	1.76	1.45	2.61	1.70	1.45
		Γ	3.83	3.83	3.46	3.06	3.26	3.25	3.57	3.33	3.03
		R	3.83	3.81	3.63	3.04	3.59	3.48	3.55	3.69	3.42
	SE-in	Γ	5.42	5.28	5.21	4.10	4.25	4.47	5.00	4.73	5.03
		R	6.35	6.14	5.89	4.45	4.90	4.91	5.65	5.57	5.34
	ME-in	Γ	5.22	5.10	4.99	4.05	4.25	4.39	4.75	4.59	4.87
		R	5.53	5.50	5.33	4.23	4.69	4.55	5.09	5.08	5.02
	ME-eff	Γ	5.31	5.45	5.17	3.87	4.26	4.04	4.89	4.32	4.75
		R	5.33	5.29	5.06	3.83	4.30	4.17	4.79	4.89	4.73
	DM	Γ	1.10	3.04	4.26	1.11	2.65	4.15	1.00	2.72	4.01
		R	3.46	3.75	3.91	3.67	4.23	4.62	3.67	3.79	4.20
CLR	Asymp.		12.05	12.02	11.66	9.64	10.50	9.69	11.22	10.96	10.77
	EE	M	2.38	1.89	1.67	2.05	1.75	1.45	2.61	1.69	1.45
		Γ	3.86	3.83	3.45	3.03	3.26	3.24	3.54	3.33	3.03
		R	3.82	3.81	3.64	3.05	3.59	3.48	3.57	3.68	3.42
	SE-in	Γ	5.42	5.28	5.21	4.12	4.25	4.45	5.01	4.71	5.02
		R	6.35	6.14	5.88	4.46	4.90	4.88	5.63	5.54	5.34
	SE-eff	Γ	5.41	5.41	5.11	4.01	4.30	4.17	4.92	4.54	4.87
		R	6.03	5.87	5.43	4.12	4.59	4.34	5.17	5.16	4.98
$F_{1^{st}}$	Asymp.		100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
F_{eff}	Asymp.		96.57	99.72	99.94	92.38	97.71	98.84	93.43	97.69	98.61
F^c_{eff}	Asymp.		95.67	99.57	99.93	91.42	97.35	98.64	92.70	97.51	98.41

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters and 20 observations per cluster. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for the relative bias.

Table S.4: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 0.1$, endogeneity degree $\rho = 0.95$, varying error distribution

			$\mathcal{N}$ errors			χ^2 errors			t errors		
Stat.	Method	$\kappa \rightarrow$	0	1	2	0	1	2	0	1	2
Wald	Asymp.		93.67	92.68	91.84	91.54	89.37	89.49	91.44	90.27	89.25
	ME-IV	Γ	82.13	81.19	78.84	76.77	74.75	75.11	77.96	76.59	75.32
		R	82.67	80.93	78.22	76.79	74.68	74.05	78.51	76.76	74.40
	ME-eff	Γ	44.02	43.14	41.90	39.35	39.03	37.74	40.65	40.77	38.65
		R	31.54	30.57	29.65	29.65	27.59	26.63	30.47	29.68	28.84
	Pairs		71.61	68.60	64.71	57.59	55.49	53.68	63.92	60.82	57.16
AR	Asymp.		17.08	16.99	16.95	13.33	13.75	13.41	15.56	15.89	15.22
	EE	M	4.42	4.30	4.06	2.93	3.10	2.65	3.51	3.10	2.90
		Γ	6.30	6.23	6.17	4.75	5.24	4.68	5.51	5.54	5.46
		R	4.46	4.48	4.43	3.52	4.34	3.75	4.02	4.16	4.06
	SE-in	Γ	6.89	6.90	6.82	5.00	5.48	5.06	6.57	6.60	6.29
		R	5.39	5.28	5.24	3.87	4.44	3.89	4.99	5.20	5.04
	SE-eff	Γ	6.58	6.92	6.77	5.38	5.72	5.34	6.63	6.45	6.42
		R	5.07	5.00	5.00	3.98	4.41	4.08	4.67	4.72	4.63
	DM	Γ	2.69	1.76	0.54	2.25	1.40	0.66	2.26	1.54	0.73
		R	5.59	5.63	5.71	5.05	5.30	5.43	5.45	5.71	5.66
KLM	Asymp.		34.78	34.80	33.43	28.63	28.06	27.51	29.93	30.24	29.79
	EE	M	9.51	9.05	8.16	6.40	6.03	5.24	7.86	7.10	6.24
		Γ	15.83	15.92	14.87	11.69	11.58	11.41	13.15	13.00	12.56
		R	15.35	15.47	14.60	11.40	11.62	10.99	12.52	12.65	12.15
	SE-in	Γ	17.05	17.63	16.30	12.68	12.73	12.12	14.62	14.43	14.08
		R	17.84	18.35	17.36	13.32	13.14	12.37	14.95	15.00	14.31
	ME-in	Γ	16.68	16.93	15.83	12.76	12.37	12.05	14.13	14.13	13.87
		R	13.63	14.02	13.04	10.30	10.24	9.65	11.84	12.03	11.65
	ME-eff	Γ	16.85	17.20	16.23	12.82	12.82	12.45	14.35	14.42	14.08
		R	11.37	11.56	11.27	9.43	9.54	9.14	10.10	10.44	10.15
	DM	Γ	2.78	3.84	2.84	1.94	2.75	2.72	2.08	2.89	3.00
		R	4.10	4.25	4.19	3.25	3.76	3.88	3.59	3.93	4.03
CLR	Asymp.		33.72	34.05	32.90	27.65	27.48	26.97	29.37	29.62	29.19
	EE	M	8.77	8.38	7.47	5.77	5.47	4.63	7.04	6.49	5.65
		Γ	14.91	15.04	14.15	10.81	10.89	10.78	12.17	12.36	11.91
		R	13.95	14.08	13.52	10.35	10.87	10.34	11.39	11.82	11.52
	SE-in	Γ	16.29	16.57	15.37	11.75	11.99	11.73	13.81	13.90	13.67
		R	16.43	17.03	16.11	12.29	12.28	11.58	13.92	14.35	13.64
	SE-eff	Γ	16.18	16.47	15.14	12.03	12.39	11.71	13.69	13.73	13.38
		R	15.64	16.29	15.29	12.06	12.01	11.42	13.13	13.63	12.93
$\mathrm{F}_{1^{st}}$	Asymp.		76.29	79.69	80.83	76.63	78.73	81.43	75.11	78.39	81.34
F_{eff}	Asymp.		0.29	0.41	0.64	0.35	0.53	0.82	0.35	0.58	0.98
F^{c}_{eff}	Asymp.		0.21	0.29	0.42	0.25	0.41	0.69	0.27	0.45	0.77

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters and 20 observations per cluster. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for the relative bias.

Table S.5: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 18$, endogeneity degree $\rho = 0.95$, normal random errors, varying number of excluded instruments

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$k_z \rightarrow$	2	5	10	15	2	5	10	15	2	5	10	15
Wald	Asymp.		16.16	33.70	24.18	29.50	17.94	41.88	24.24	32.80	21.34	45.73	24.47	36.68
	ME-IV	Γ	13.86	27.46	17.81	20.26	15.20	34.65	17.80	25.46	16.87	38.03	17.52	29.15
		R	13.93	30.01	15.39	22.80	15.38	37.60	15.81	27.59	17.10	41.44	15.67	31.79
	ME-eff	Γ	7.27	0.52	14.41	3.13	7.44	0.77	14.40	3.81	7.67	1.19	12.91	4.90
		R	5.58	1.42	7.81	5.72	5.70	2.06	8.06	6.00	5.65	2.40	7.66	6.42
	Pairs		8.33	10.58	8.90	13.31	9.74	18.00	9.82	18.12	11.54	22.26	10.49	21.73
AR	Asymp.		6.86	17.08	60.93	96.98	6.59	16.99	61.03	97.03	6.19	16.95	60.44	96.99
	EE	M	3.21	4.42	53.67	99.01	3.12	4.30	54.40	99.13	2.51	4.06	53.46	99.05
		Γ	5.79	6.30	7.37	5.99	5.29	6.23	7.21	6.44	5.15	6.17	7.83	5.86
		R	4.70	4.46	4.30	2.64	4.70	4.48	4.25	2.87	4.70	4.43	4.47	2.53
	SE-in	Γ	5.92	6.89	8.52	7.30	5.81	6.90	8.30	7.64	5.60	6.82	8.66	6.85
		R	5.39	5.39	5.37	5.85	5.30	5.28	5.27	6.00	5.26	5.24	5.47	5.51
	SE-eff	Γ	5.96	6.58	8.78	7.39	5.69	6.92	8.64	7.86	5.64	6.77	8.95	7.09
		R	5.38	5.07	5.02	5.69	5.21	5.00	4.91	5.66	5.04	5.00	5.07	5.36
	DM	Γ	2.64	2.69	2.90	0.14	3.62	1.76	1.01	0.01	3.19	0.54	0.06	0.00
		R	5.44	5.59	6.50	8.15	5.53	5.63	6.46	8.68	5.63	5.71	6.85	8.85
KLM	Asymp.		7.74	12.13	37.00	67.60	7.28	12.12	37.72	68.18	6.81	11.81	36.44	66.86
	EE	M	4.43	2.38	24.00	66.63	3.75	1.89	23.83	65.93	3.26	1.68	22.26	64.68
		Γ	5.24	3.83	6.60	5.81	4.83	3.83	6.28	5.77	4.59	3.46	6.19	5.75
		R	4.86	3.83	5.29	3.46	4.70	3.81	4.91	3.59	4.60	3.63	5.11	3.28
	SE-in	Γ	5.60	5.42	7.35	7.01	5.11	5.28	7.02	7.13	4.78	5.21	7.14	6.69
		R	5.34	6.35	6.61	6.75	5.30	6.14	6.32	6.51	5.28	5.89	6.39	6.20
	ME-in	Γ	5.56	5.22	7.42	7.01	5.24	5.10	7.13	7.13	4.95	4.99	7.21	6.73
		R	5.16	5.53	6.41	6.63	5.33	5.50	6.25	6.58	5.25	5.33	6.26	6.21
	ME-eff	Γ	5.34	5.31	7.36	6.68	5.18	5.45	7.25	7.04	4.89	5.17	7.30	6.41
		R	5.06	5.33	5.65	5.63	5.22	5.29	5.58	5.59	5.03	5.06	5.61	5.45
	DM	Γ	3.69	1.10	2.05	2.09	4.67	3.04	3.99	4.19	5.63	4.26	5.45	3.72
		R	5.03	3.46	4.28	4.56	4.90	3.75	4.53	4.62	5.13	3.91	4.91	4.84
CLR	Asymp.		7.80	12.05	63.58	52.42	7.32	12.02	63.25	52.14	6.89	11.66	60.27	51.62
	EE	M	4.42	2.38	12.47	4.53	3.77	1.89	11.77	3.50	3.23	1.67	10.10	2.67
		Γ	5.18	3.86	2.58	0.45	4.94	3.83	2.53	0.51	4.54	3.45	2.26	0.22
		R	4.83	3.82	1.82	0.31	4.71	3.81	1.86	0.34	4.62	3.64	1.72	0.18
	SE-in	Γ	5.63	5.42	3.01	0.49	5.13	5.28	2.80	0.56	4.81	5.21	2.61	0.28
		R	5.27	6.35	2.31	0.50	5.25	6.14	2.21	0.45	5.26	5.88	2.06	0.23
	SE-eff	Γ	5.44	5.41	3.10	0.53	5.24	5.41	2.84	0.57	4.90	5.11	2.55	0.25
		R	5.13	6.03	2.16	0.39	5.21	5.87	2.12	0.40	5.10	5.43	1.86	0.23
$F_{1^{st}}$	Asymp.		99.97	100.00	100.00	100.00	99.99	100.00	100.00	100.00	99.97	100.00	100.00	100.00
F_{eff}	Asymp.		95.01	96.57	99.22	100.00	96.22	99.72	99.64	100.00	97.57	99.94	99.84	100.00
F_{eff}^{c}	Asymp.		83.49	95.67	99.09	100.00	88.21	99.57	99.61	100.00	93.36	99.93	99.82	100.00

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters and 20 observations per cluster. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for bias.

Table S.6: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 0.1$, endogeneity degree $\rho = 0.95$, normal random errors, varying number of excluded instruments

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$k_z \rightarrow$	2	5	10	15	2	5	10	15	2	5	10	15
Wald	Asymp.		69.20	93.67	99.54	99.95	69.93	92.68	99.26	99.89	69.76	91.84	99.29	99.89
	ME-IV	Γ	55.34	82.13	96.20	99.37	55.99	81.19	95.85	98.79	55.80	78.84	95.02	98.39
		R	56.97	82.67	96.02	98.68	57.94	80.93	95.54	97.70	57.63	78.22	94.53	96.77
	ME-eff	Γ	25.27	44.02	81.01	99.03	26.64	43.14	80.61	98.37	27.29	41.90	79.10	97.88
		R	19.05	31.54	71.45	97.13	20.03	30.57	70.63	95.98	21.34	29.65	69.06	95.09
	Pairs		49.43	71.61	90.35	94.82	50.45	68.60	89.29	91.45	49.82	64.71	87.43	88.62
AR	Asymp.		6.86	17.08	60.93	96.98	6.59	16.99	61.03	97.03	6.19	16.95	60.44	96.99
	EE	M	3.21	4.42	53.67	99.01	3.12	4.30	54.40	99.13	2.51	4.06	53.46	99.05
		Γ	5.79	6.30	7.37	5.99	5.29	6.23	7.21	6.44	5.15	6.17	7.83	5.86
		R	4.70	4.46	4.30	2.64	4.70	4.48	4.25	2.87	4.70	4.43	4.47	2.53
	SE-in	Γ	5.92	6.89	8.52	7.30	5.81	6.90	8.30	7.64	5.60	6.82	8.66	6.85
		R	5.39	5.39	5.37	5.85	5.30	5.28	5.27	6.00	5.26	5.24	5.47	5.51
	SE-eff	Γ	5.96	6.58	8.78	7.39	5.69	6.92	8.64	7.86	5.64	6.77	8.95	7.09
		R	5.38	5.07	5.02	5.69	5.21	5.00	4.91	5.66	5.04	5.00	5.07	5.36
	DM	Γ	2.64	2.69	2.90	0.14	3.62	1.76	1.01	0.01	3.19	0.54	0.06	0.00
		R	5.44	5.59	6.50	8.15	5.53	5.63	6.46	8.68	5.63	5.71	6.85	8.85
KLM	Asymp.		12.89	34.78	69.18	86.06	12.90	34.80	69.80	85.94	12.33	33.43	69.47	84.82
	EE	M	5.80	9.51	53.48	84.45	5.35	9.05	53.64	84.27	4.56	8.16	53.02	83.35
		Γ	9.43	15.83	23.43	12.68	9.01	15.92	22.97	13.10	9.06	14.87	23.02	12.45
		R	9.04	15.35	20.35	6.21	8.91	15.47	19.71	6.65	8.92	14.60	20.11	6.37
	SE-in	Γ	9.68	17.05	24.48	14.27	9.37	17.63	24.35	14.77	9.12	16.30	24.62	14.01
		R	9.33	17.84	22.97	13.65	9.69	18.35	22.69	14.34	9.45	17.36	23.18	13.56
	ME-in	Γ	9.83	16.68	25.36	13.76	9.60	16.93	25.14	14.39	9.48	15.83	25.07	13.67
		R	8.27	13.63	18.68	12.23	8.79	14.02	18.40	12.87	8.63	13.04	18.49	12.22
	ME-eff	Γ	9.59	16.85	25.03	13.77	9.48	17.20	24.90	14.13	9.38	16.23	24.97	13.94
		R	7.84	11.37	16.43	11.12	8.02	11.56	16.04	11.25	8.05	11.27	16.41	11.28
	DM	Γ	3.34	2.78	3.44	2.64	4.11	3.84	3.84	2.89	4.24	2.84	3.02	1.22
		R	5.03	4.10	4.51	4.76	4.99	4.25	4.85	5.01	4.89	4.19	4.82	5.35
CLR	Asymp.		11.33	33.72	82.70	80.48	11.35	34.05	83.98	79.24	10.94	32.90	83.28	77.99
	EE	M	4.61	8.77	30.15	45.46	4.21	8.38	29.90	43.99	3.72	7.47	30.41	42.05
		Γ	8.46	14.91	11.25	4.79	8.06	15.04	10.93	4.89	8.13	14.15	11.60	4.19
		R	7.76	13.95	9.54	2.77	7.77	14.08	8.97	2.75	8.05	13.52	9.98	2.50
	SE-in	Γ	8.60	16.29	11.71	5.30	8.52	16.57	11.66	5.49	8.36	15.37	12.21	5.07
		R	8.29	16.43	10.67	5.00	8.63	17.03	10.58	4.90	8.75	16.11	11.28	4.51
	SE-eff	Γ	8.55	16.18	11.85	5.41	8.55	16.47	11.79	5.34	8.30	15.14	12.15	5.02
		R	8.11	15.64	9.97	4.49	8.29	16.29	9.66	4.45	8.41	15.29	10.48	4.04
$F_{1^{st}}$	Asymp.		31.83	76.29	93.27	100.00	34.27	79.69	93.40	100.00	37.32	80.83	93.66	100.00
F_{eff}	Asymp.		2.73	0.29	0.48	20.19	3.12	0.41	0.59	21.84	4.00	0.64	0.61	23.47
F^c_{eff}	Asymp.		0.35	0.21	0.47	20.11	0.58	0.29	0.55	21.80	0.83	0.42	0.57	23.42

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters and 20 observations per cluster. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for bias.

Table S.7: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 18$, endogeneity degree $\rho = 0.95$, normal random errors, varying number of clusters

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$G \rightarrow$	10	20	40	80	10	20	40	80	10	20	40	80
Wald	Asymp.		38.18	33.70	28.08	17.13	43.39	41.88	38.15	24.86	46.77	45.73	44.22	27.98
	ME-IV	Γ	24.99	27.46	26.62	16.22	28.91	34.65	35.39	21.29	33.13	38.03	39.25	21.57
		R	26.41	30.01	28.64	15.42	30.20	37.60	38.54	21.17	34.04	41.44	43.03	21.03
	ME-eff	Γ	0.21	0.52	1.54	9.09	0.27	0.77	2.87	9.75	0.37	1.19	4.09	10.14
		R	1.52	1.42	2.10	6.11	1.73	2.06	2.91	6.41	2.01	2.40	3.50	6.64
	Pairs		8.04	10.58	10.75	9.50	10.76	18.00	22.03	15.56	14.49	22.26	27.24	16.82
AR	Asymp.		47.50	17.08	8.32	5.63	47.32	16.99	7.70	5.18	46.91	16.95	8.12	4.88
	EE	M	53.54	4.42	1.60	1.17	53.44	4.30	1.03	0.86	53.03	4.06	1.04	0.77
		Γ	6.05	6.30	6.22	6.08	5.87	6.23	5.93	5.86	5.72	6.17	5.92	5.46
		R	2.94	4.46	4.79	5.02	2.97	4.48	4.64	4.95	2.72	4.43	4.65	4.76
	SE-in	Γ	7.62	6.89	6.18	5.95	7.17	6.90	5.90	5.84	7.48	6.82	5.78	5.42
		R	6.65	5.39	4.99	5.11	6.70	5.28	4.70	5.03	6.42	5.24	4.69	4.89
	SE-eff	Γ	7.41	6.58	6.18	5.91	6.81	6.92	6.00	5.89	7.09	6.77	5.89	5.47
		R	5.40	5.07	4.93	5.17	5.22	5.00	4.67	5.20	5.11	5.00	4.68	4.85
	DM	Γ	1.43	2.69	2.72	2.54	0.59	1.76	2.00	2.37	0.10	0.54	1.06	1.51
		R	7.93	5.59	5.05	5.17	7.98	5.63	5.14	5.40	8.42	5.71	5.25	5.45
KLM	Asymp.		37.72	12.13	6.54	6.52	38.36	12.12	5.42	6.28	37.49	11.81	5.71	6.28
	EE	M	22.72	2.38	1.73	2.24	22.44	1.89	1.15	1.33	21.79	1.68	1.04	1.08
		Γ	4.38	3.83	4.45	5.38	4.48	3.83	3.54	5.07	4.35	3.46	3.85	5.11
		R	2.89	3.83	4.97	5.35	2.80	3.81	4.26	5.23	2.64	3.63	4.55	5.29
	SE-in	Γ	6.38	5.42	5.03	5.23	6.39	5.28	4.35	4.97	6.09	5.21	4.65	5.04
		R	6.95	6.35	6.07	5.47	6.89	6.14	5.61	5.56	6.86	5.89	6.01	5.58
	ME-in	Γ	6.32	5.22	4.65	5.20	6.27	5.10	4.12	4.89	6.09	4.99	4.45	4.94
		R	6.84	5.53	4.97	5.19	6.84	5.50	4.64	5.27	6.82	5.33	5.08	5.38
	ME-eff	Γ	5.86	5.31	4.94	5.08	5.59	5.45	4.32	5.00	5.46	5.17	4.68	4.91
		R	5.43	5.33	5.30	5.32	5.27	5.29	4.73	5.29	5.20	5.06	5.33	5.32
	DM	Γ	0.05	1.10	2.35	3.72	0.40	3.04	3.94	5.33	2.07	4.26	5.38	5.51
		R	0.91	3.46	4.84	5.13	0.89	3.75	4.64	5.38	1.24	3.91	5.23	5.45
CLR	Asymp.		40.12	12.05	6.51	6.46	40.74	12.02	5.34	6.27	39.82	11.66	5.69	6.23
	EE	M	22.29	2.38	1.69	2.18	22.04	1.89	1.12	1.31	21.53	1.67	1.04	1.08
		Γ	4.35	3.86	4.49	5.39	4.46	3.83	3.52	5.06	4.26	3.45	3.86	5.09
		R	2.81	3.82	4.90	5.35	2.77	3.81	4.26	5.24	2.59	3.64	4.54	5.29
	SE-in	Γ	6.27	5.42	5.03	5.25	6.23	5.28	4.33	5.00	5.99	5.21	4.64	5.00
		R	6.66	6.35	6.06	5.48	6.67	6.14	5.60	5.53	6.79	5.88	6.00	5.59
	SE-eff	Γ	5.86	5.41	5.02	5.22	5.54	5.41	4.36	4.98	5.44	5.11	4.68	4.91
		R	5.36	6.03	6.00	5.58	5.31	5.87	5.41	5.39	5.24	5.43	5.87	5.55
$F_{1^{st}}$	Asymp.		100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
F_{eff}	Asymp.		98.83	96.57	94.48	94.22	99.49	99.72	99.88	99.62	99.86	99.94	99.98	99.54
F^c_{eff}	Asymp.		98.80	95.67	91.84	89.11	99.47	99.57	99.81	99.33	99.86	99.93	99.96	99.29

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is $G \times 20$ observations with 20 observations clusters. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for bias.

Table S.8: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 0.1$, endogeneity degree $\rho = 0.95$, normal random errors, varying number of clusters

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$G \rightarrow$	10	20	40	80	10	20	40	80	10	20	40	80
Wald	Asymp.		94.84	93.67	93.25	92.82	96.20	94.97	94.28	95.46	99.20	99.51	99.51	99.83
	ME-IV	Γ	82.04	82.13	81.98	81.69	85.11	85.97	85.02	86.67	95.29	97.46	98.27	98.68
		R	80.53	82.67	82.48	82.16	83.51	85.31	85.02	87.07	94.13	97.38	98.36	98.72
	ME-eff	Γ	62.27	44.02	32.43	29.79	66.33	51.54	43.94	45.92	85.85	87.66	90.71	92.28
		R	48.64	31.54	22.80	22.44	52.30	36.63	33.18	38.60	73.02	80.51	88.54	91.72
	Pairs		61.76	71.61	75.57	76.84	64.51	74.57	76.97	81.44	80.85	94.15	96.85	98.09
AR	Asymp.		47.50	17.08	8.32	5.63	47.68	16.74	7.48	4.99	43.03	13.35	5.39	3.23
	EE	M	53.54	4.42	1.60	1.17	53.96	3.88	0.65	0.51	48.87	1.77	0.07	0.00
		Γ	6.05	6.30	6.22	6.08	5.71	6.56	5.86	5.62	5.00	4.72	4.25	4.62
		R	2.94	4.46	4.79	5.02	2.94	4.79	4.50	5.10	2.44	3.92	3.57	4.50
	SE-in	Γ	7.62	6.89	6.18	5.95	7.09	7.06	5.76	5.60	6.37	6.69	5.44	5.56
		R	6.65	5.39	4.99	5.11	6.49	5.46	4.77	5.05	5.93	5.09	4.35	5.00
	SE-eff	Γ	7.41	6.58	6.18	5.91	6.83	7.05	5.70	5.59	5.90	6.64	5.41	5.66
		R	5.40	5.07	4.93	5.17	5.46	5.16	4.70	5.02	5.08	5.39	4.79	5.28
	DM	Γ	1.43	2.69	2.72	2.54	0.26	0.81	1.33	2.08	0.92	5.91	8.18	8.70
		R	7.93	5.59	5.05	5.17	9.19	6.42	5.50	5.60	15.62	14.29	11.51	12.11
KLM	Asymp.		58.32	34.78	23.48	21.07	59.94	37.02	25.56	23.55	65.05	36.13	20.85	21.24
	EE	M	44.22	9.51	5.64	6.02	44.96	8.56	2.94	2.90	47.91	4.13	0.39	0.28
		Γ	10.25	15.83	16.74	17.76	10.31	17.25	18.22	20.67	9.44	14.56	13.82	18.62
		R	5.82	15.35	16.94	18.28	6.15	16.97	18.32	20.91	5.59	15.20	13.85	18.52
	SE-in	Γ	12.39	17.05	16.70	17.23	12.90	18.98	18.71	20.02	12.48	20.26	17.29	20.92
		R	13.43	17.84	17.77	18.31	13.71	19.45	19.14	20.83	14.00	18.98	14.98	19.61
	ME-in	Γ	12.61	16.68	15.98	17.27	12.96	18.07	17.92	19.74	12.55	19.22	16.23	20.43
		R	12.49	13.63	12.83	13.44	12.81	14.91	13.74	15.54	12.64	14.69	11.25	15.25
	ME-eff	Γ	12.02	16.85	15.45	15.45	11.99	18.51	17.23	17.90	11.55	19.35	15.82	19.30
		R	9.91	11.37	10.19	10.22	9.86	12.33	10.93	11.85	10.83	13.28	10.01	13.20
	DM	Γ	2.29	2.78	2.77	3.35	3.07	3.89	3.94	4.87	5.14	7.43	8.99	4.85
		R	4.11	4.10	4.16	4.57	4.77	5.02	5.15	5.43	8.17	9.86	9.62	6.15
CLR	Asymp.		72.08	33.72	20.64	16.37	73.22	36.19	23.17	19.92	70.09	35.92	20.06	20.09
	EE	M	18.29	8.77	3.88	3.72	17.80	7.87	1.81	1.67	14.43	3.88	0.19	0.06
		Γ	4.27	14.91	14.87	14.91	4.43	16.31	16.59	17.96	3.11	14.21	13.47	17.88
		R	2.57	13.95	14.37	14.99	2.75	15.73	16.45	18.41	1.93	14.87	13.47	17.66
	SE-in	Γ	5.13	16.29	14.59	14.53	5.14	17.79	17.03	17.72	3.89	20.03	16.67	20.14
		R	5.14	16.43	15.25	14.89	4.82	18.27	17.37	18.43	3.78	18.58	14.42	18.66
	SE-eff	Γ	4.88	16.18	14.54	14.50	4.66	17.73	17.01	17.76	3.44	19.59	16.40	20.28
		R	3.80	15.64	15.04	14.93	3.63	17.50	17.04	18.47	3.20	18.21	14.53	18.97
$F_{1^{st}}$	Asymp.		95.87	76.29	53.12	36.94	96.54	82.94	68.10	52.62	98.50	95.88	94.13	53.87
F_{eff}	Asymp.		9.50	0.29	0.00	0.00	11.13	0.87	0.27	0.07	32.01	37.43	18.73	0.24
F^c_{eff}	Asymp.		9.15	0.21	0.00	0.00	10.72	0.72	0.13	0.02	31.35	34.15	13.82	0.12

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is $G \times 20$ observations with 20 observations per cluster. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for bias.

Table S.9: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 18$, normal random errors, varying degree of endogeneity

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$\rho \rightarrow$	0.20	0.50	0.70	0.95	0.20	0.50	0.70	0.95	0.20	0.50	0.70	0.95
Wald	Asymp.		33.89	33.90	33.85	33.70	42.93	42.63	42.46	41.88	47.83	47.37	46.60	45.73
	ME-IV	Γ	31.25	30.76	29.55	27.46	37.48	36.77	35.85	34.65	41.56	40.30	39.58	38.03
		R	34.12	33.23	31.97	30.01	41.22	40.36	39.39	37.60	45.29	44.23	43.07	41.44
	ME-eff	Γ	0.67	0.65	0.70	0.52	0.85	0.92	0.84	0.77	0.97	1.14	1.28	1.19
		R	1.84	1.75	1.65	1.42	2.15	2.26	2.30	2.06	2.19	2.23	2.40	2.40
	Pairs		20.70	18.14	15.24	10.58	27.85	25.37	22.53	18.00	31.73	29.46	27.01	22.26
AR	Asymp.		17.08	17.08	17.08	17.08	16.99	16.99	16.99	16.99	16.95	16.95	16.95	16.95
	EE	M	4.76	4.76	4.76	4.76	4.56	4.56	4.56	4.56	4.22	4.22	4.22	4.22
		Γ	6.15	6.15	6.15	6.15	6.25	6.25	6.25	6.25	6.12	6.12	6.12	6.12
		R	4.32	4.32	4.32	4.32	4.14	4.14	4.14	4.14	4.25	4.25	4.25	4.25
	SE-in	Γ	6.89	6.89	6.89	6.89	6.90	6.90	6.90	6.90	6.82	6.82	6.82	6.82
		R	5.39	5.39	5.39	5.39	5.28	5.28	5.28	5.28	5.24	5.24	5.24	5.24
	SE-eff	Γ	6.58	6.58	6.58	6.58	6.92	6.92	6.92	6.92	6.77	6.77	6.77	6.77
		R	5.07	5.07	5.07	5.07	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
	DM	Γ	2.69	2.69	2.69	2.69	1.76	1.76	1.76	1.76	0.54	0.54	0.54	0.54
		R	5.59	5.59	5.59	5.59	5.63	5.63	5.63	5.63	5.71	5.71	5.71	5.71
KLM	Asymp.		11.98	11.93	12.14	12.13	11.61	11.60	11.92	12.12	11.78	11.85	11.88	11.81
	EE	M	2.52	2.45	2.44	2.46	1.97	2.02	2.03	2.14	1.77	1.77	1.79	1.81
		Γ	3.35	3.51	3.65	3.62	3.29	3.43	3.58	3.74	3.17	3.34	3.39	3.46
		R	3.32	3.35	3.42	3.60	3.30	3.47	3.54	3.65	3.08	3.18	3.25	3.42
	SE-in	Γ	4.85	5.10	5.33	5.42	4.75	4.76	5.02	5.28	4.82	4.97	5.10	5.21
		R	5.44	5.63	5.78	6.35	5.40	5.60	5.74	6.14	5.27	5.25	5.46	5.89
	ME-in	Γ	4.79	5.01	5.14	5.22	4.89	4.83	4.83	5.10	4.88	4.85	4.94	4.99
		R	5.03	5.27	5.36	5.53	5.17	5.31	5.35	5.50	5.10	5.17	5.12	5.33
	ME-eff	Γ	4.66	4.88	5.05	5.31	4.57	4.87	4.97	5.45	4.60	4.82	4.75	5.17
		R	4.59	4.75	4.90	5.33	4.64	4.87	4.87	5.29	4.36	4.45	4.60	5.06
	DM	Γ	1.07	1.04	1.01	1.10	2.90	2.86	2.84	3.04	3.94	3.89	4.05	4.26
		R	3.14	3.18	3.24	3.46	3.44	3.47	3.49	3.75	3.69	3.70	3.82	3.91
CLR	Asymp.		11.92	11.85	12.13	12.05	11.56	11.54	11.86	12.02	11.75	11.79	11.74	11.66
	EE	M	2.55	2.47	2.44	2.44	1.97	2.02	2.04	2.14	1.79	1.77	1.79	1.81
		Γ	3.36	3.49	3.63	3.63	3.29	3.43	3.58	3.73	3.17	3.34	3.36	3.47
		R	3.34	3.36	3.42	3.61	3.31	3.47	3.54	3.65	3.08	3.19	3.25	3.45
	SE-in	Γ	4.87	5.10	5.33	5.42	4.76	4.75	5.00	5.28	4.82	4.95	5.10	5.21
		R	5.42	5.62	5.75	6.35	5.40	5.60	5.74	6.14	5.28	5.25	5.43	5.88
	SE-eff	Γ	4.84	5.01	5.15	5.41	4.78	4.87	5.04	5.41	4.65	4.83	4.93	5.11
		R	4.92	5.28	5.42	6.03	5.04	5.20	5.38	5.87	4.81	4.88	4.93	5.43
$F_{1^{st}}$	Asymp.		100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
F_{eff}	Asymp.		96.27	96.37	96.50	96.57	99.67	99.64	99.64	99.72	99.96	99.91	99.91	99.94
F^{c}_{eff}	Asymp.		95.29	95.45	95.66	95.67	99.61	99.54	99.49	99.57	99.92	99.87	99.87	99.93

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters and 20 observations per cluster. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for bias.

Table S.10: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with first-stage F-test noncentrality parameter $\mu_{k_z} = 0.1$, normal random errors, varying degree of endogeneity

			$\kappa = 0$				$\kappa = 1$				$\kappa = 2$			
Stat.	Method	$\rho \rightarrow$	0.20	0.50	0.70	0.95	0.20	0.50	0.70	0.95	0.20	0.50	0.70	0.95
Wald	Asymp.		12.90	33.21	58.29	93.67	14.48	36.18	60.26	92.68	16.04	36.36	60.04	91.84
	ME-IV	Γ	6.70	21.60	42.89	82.13	7.95	23.66	45.40	81.19	8.61	24.06	44.83	78.84
		R	7.26	23.25	45.24	82.67	8.66	25.13	47.47	80.93	9.07	25.57	46.57	78.22
	ME-eff	Γ	3.37	10.70	20.41	44.02	3.51	11.46	22.06	43.14	3.78	11.57	21.72	41.90
		R	4.25	9.80	15.41	31.54	4.28	9.81	16.07	30.57	4.48	10.17	15.90	29.65
	Pairs		2.82	13.34	31.17	71.61	3.41	14.97	33.78	68.60	3.82	15.40	33.63	64.71
AR	Asymp.		17.08	17.08	17.08	17.08	16.99	16.99	16.99	16.99	16.95	16.95	16.95	16.95
	EE	M	4.76	4.76	4.76	4.76	4.56	4.56	4.56	4.56	4.22	4.22	4.22	4.22
		Γ	6.15	6.15	6.15	6.15	6.25	6.25	6.25	6.25	6.12	6.12	6.12	6.12
		R	4.32	4.32	4.32	4.32	4.14	4.14	4.14	4.14	4.25	4.25	4.25	4.25
	SE-in	Γ	6.89	6.89	6.89	6.89	6.90	6.90	6.90	6.90	6.82	6.82	6.82	6.82
		R	5.39	5.39	5.39	5.39	5.28	5.28	5.28	5.28	5.24	5.24	5.24	5.24
	SE-eff	Γ	6.58	6.58	6.58	6.58	6.92	6.92	6.92	6.92	6.77	6.77	6.77	6.77
		R	5.07	5.07	5.07	5.07	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
	DM	Γ	2.69	2.69	2.69	2.69	1.76	1.76	1.76	1.76	0.54	0.54	0.54	0.54
		R	5.59	5.59	5.59	5.59	5.63	5.63	5.63	5.63	5.71	5.71	5.71	5.71
KLM	Asymp.		15.91	18.91	22.91	34.78	15.96	18.67	22.62	34.80	15.99	18.66	22.20	33.43
	EE	M	5.70	6.23	7.18	9.85	5.25	5.87	7.06	9.38	5.20	5.47	6.22	8.45
		Γ	5.53	6.97	9.12	15.54	5.89	6.96	9.13	15.95	5.52	6.90	8.53	14.89
		R	4.72	6.31	8.03	15.07	5.16	6.50	8.56	15.09	5.01	6.17	8.02	14.53
	SE-in	Γ	5.95	7.62	9.51	17.05	6.26	7.70	9.68	17.63	6.08	7.49	9.38	16.30
		R	5.85	7.35	9.70	17.84	5.92	7.37	9.88	18.35	5.83	7.30	9.51	17.36
	ME-in	Γ	6.11	7.53	9.73	16.68	6.16	7.67	9.68	16.93	6.14	7.56	9.33	15.83
		R	5.26	6.23	8.30	13.63	5.73	6.71	8.62	14.02	5.45	6.44	8.00	13.04
	ME-eff	Γ	5.95	7.41	9.80	16.85	6.06	7.30	9.88	17.20	5.81	7.10	9.24	16.23
		R	4.97	5.88	7.34	11.37	5.31	6.29	7.91	11.56	5.12	6.13	7.29	11.27
	DM	Γ	3.81	3.79	3.69	2.78	3.55	3.50	3.67	3.84	1.81	1.88	2.01	2.84
		R	4.87	4.89	4.93	4.10	5.11	5.04	4.95	4.25	4.81	4.74	4.73	4.19
CLR	Asymp.		16.51	19.05	22.90	33.72	16.45	18.91	22.80	34.05	16.53	18.99	22.27	32.90
	EE	M	5.37	5.99	6.75	9.10	4.96	5.69	6.52	8.67	4.90	5.22	5.77	7.79
		Γ	5.45	6.67	8.67	14.63	5.84	6.94	8.83	15.11	5.63	6.78	8.14	14.16
		R	4.59	5.76	7.41	13.66	4.90	6.11	8.01	13.93	4.82	5.88	7.39	13.48
	SE-in	Γ	5.94	7.29	9.42	16.29	6.21	7.65	9.56	16.57	5.98	7.37	9.19	15.37
		R	5.57	6.98	9.11	16.43	5.70	7.23	9.42	17.03	5.76	7.02	8.95	16.11
	SE-eff	Γ	5.96	7.28	9.29	16.18	5.98	7.54	9.63	16.47	5.99	7.21	8.88	15.14
		R	5.21	6.52	8.62	15.64	5.35	6.70	8.82	16.29	5.33	6.49	8.42	15.29
$F_{1^{st}}$	Asymp.		75.73	76.27	76.27	76.29	77.59	78.66	78.99	79.69	79.07	79.77	80.24	80.83
F_{eff}	Asymp.		0.23	0.16	0.18	0.29	0.23	0.22	0.31	0.41	0.28	0.26	0.43	0.64
F^{c}_{eff}	Asymp.		0.16	0.11	0.11	0.21	0.19	0.16	0.24	0.29	0.22	0.20	0.35	0.42

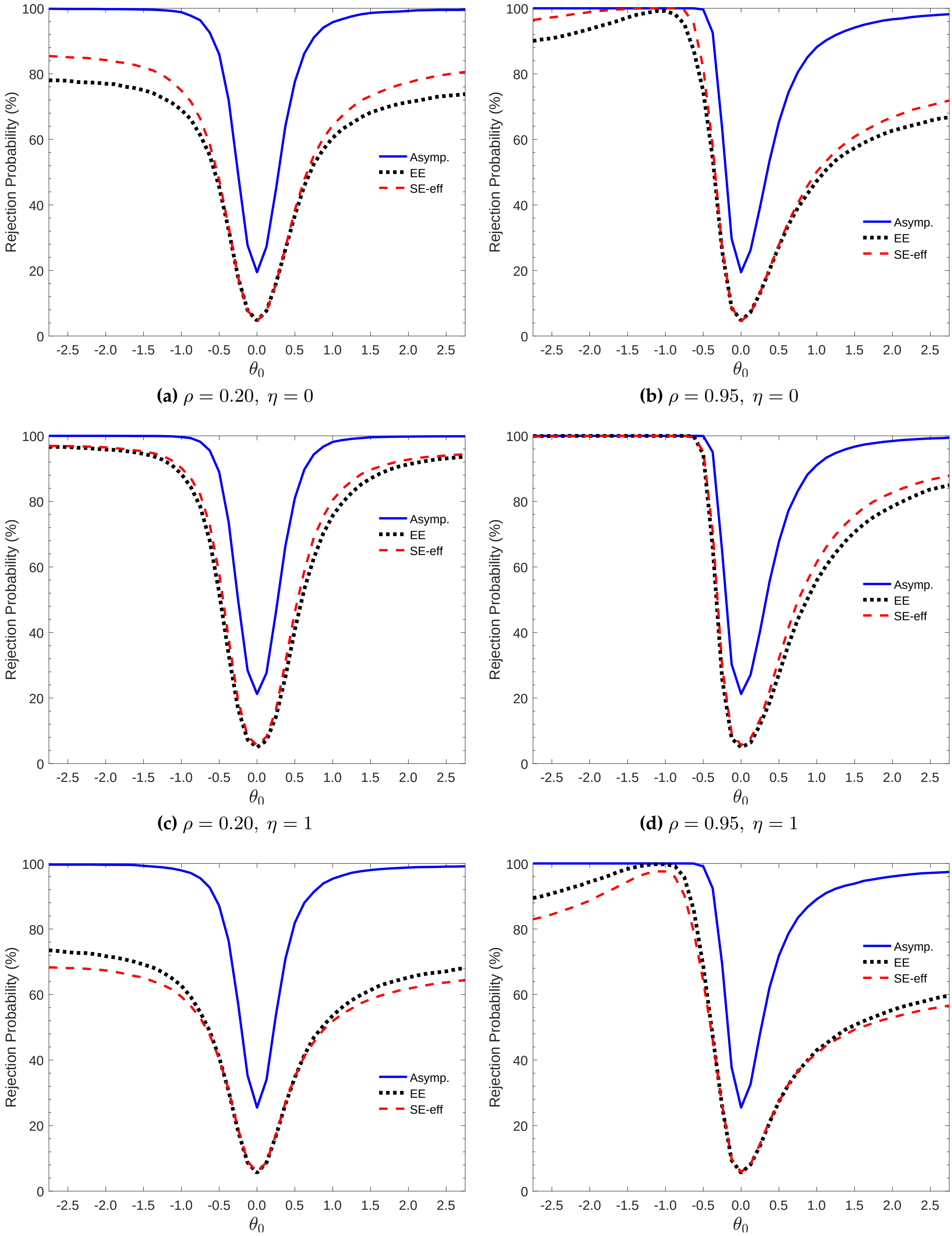
Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters and 20 observations per cluster. Number of included instruments: $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. The weights M, Γ and R correspond to the multinomial, gamma and Rademacher weights, respectively. Effective F-tests use 5% critical values under a 10% tolerance for bias.

Table S.11: Rejection percentages for testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at the 5% significance level, DGP with normal random errors, Rademacher bootstrap weights, comparison with Davidson-MacKinnon bootstraps

			$\eta = 0$			$\eta = 1$			$\eta = 2$	
	$\kappa \rightarrow$	0	1	2	0	1	2	0	1	2
Stat.	Method				Panel A: $\rho = 0.20$					
AR	Asymp.	17.08	16.74	13.35	19.01	19.52	18.55	27.07	26.52	21.79
	EE	4.46	4.79	3.92	4.32	4.39	5.29	4.20	4.68	4.01
	SE-eff	5.07	5.16	5.39	5.18	5.02	5.31	5.51	5.51	5.85
	DM	5.59	6.42	14.29	5.39	6.43	8.23	5.93	7.71	22.35
KLM	Asymp.	11.98	11.71	12.24	16.74	16.59	15.68	21.35	21.76	21.77
	EE	3.37	3.65	3.68	4.33	4.38	4.18	4.43	4.54	4.67
	ME-eff	4.59	4.63	5.04	5.03	5.01	4.84	5.21	5.32	5.94
	DM	3.14	4.01	12.32	4.30	4.74	5.46	4.37	6.47	24.78
F_{eff}	Asymp.	96.27	99.62	98.55	98.93	99.61	99.64	97.73	98.90	98.11
					Panel B: $\rho = 95$					
AR	Asymp.	17.08	16.74	13.35	19.01	19.52	18.55	27.07	26.52	21.79
	EE	4.46	4.79	3.92	4.32	4.39	5.29	4.20	4.68	4.01
	SE-eff	5.07	5.16	5.39	5.18	5.02	5.31	5.51	5.51	5.85
	DM	5.59	6.42	14.29	5.39	6.43	8.23	5.93	7.71	22.35
KLM	Asymp.	12.13	12.53	15.68	16.43	16.98	17.43	21.39	22.21	25.03
	EE	3.83	4.05	5.11	4.42	4.60	5.05	4.66	4.66	5.56
	ME-eff	5.33	5.42	6.48	5.22	5.25	5.59	5.46	5.46	7.05
	DM	3.46	4.23	11.77	4.28	4.68	5.71	4.51	6.40	25.91
F_{eff}	Asymp.	96.57	98.69	68.70	98.95	98.85	81.19	97.82	98.30	75.97

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 199 bootstrap replications for each simulation. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Noncentrality parameter of first-stage F-test: $\mu_{k_z} = 18$. Within cluster error correlation: $\phi = 0.5$. Degree of endogeneity: $\rho = 0.95$. Skedastic function: $\breve{f}(\mathbf{z}_{1,g}, \kappa) = (\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$. Effective F-test uses 5% critical values under a 10% tolerance for the relative bias.

Figure S.1: Power comparisons of bootstrapped AR tests with noncentrality parameter of first-stage F-test at $\mu_{k_z} = 18$ and residual heterogeneity at $\kappa = 0$.



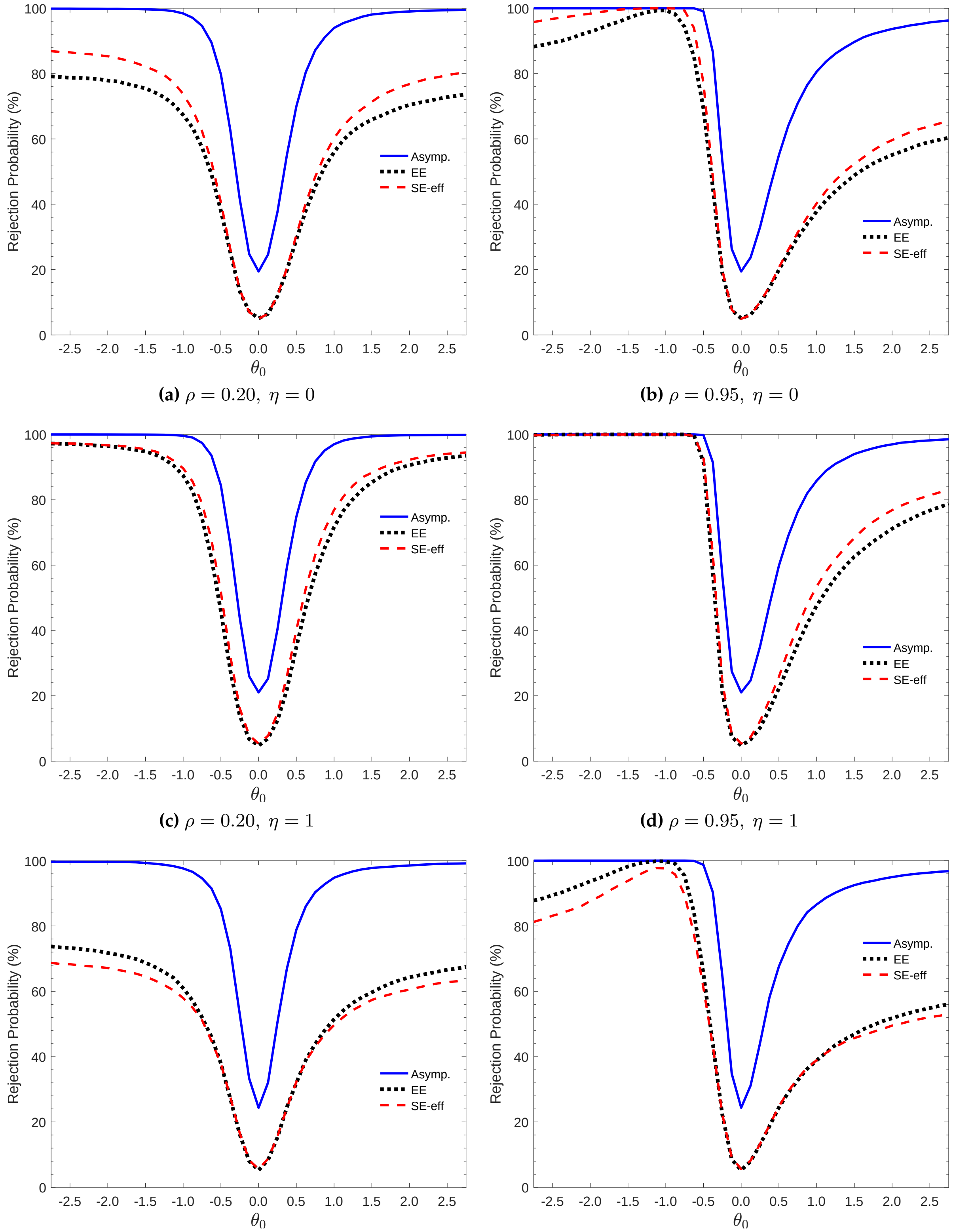
(a) $\rho = 0.20,\ \eta = 0$ **(b)** $\rho = 0.95,\ \eta = 0$

(c) $\rho = 0.20,\ \eta = 1$ **(d)** $\rho = 0.95,\ \eta = 1$

(e) $\rho = 0.20,\ \eta = 2$ **(f)** $\rho = 0.95,\ \eta = 2$

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 999 bootstrap replications for each simulation using Rademacher weights and a DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$ evaluated at $\kappa = 0$.

Figure S.2: Power comparisons of bootstrapped AR tests with noncentrality parameter of first-stage F-test at $\mu_{k_z} = 18$ and residual heterogeneity at $\kappa = 1$.



(a) $\rho = 0.20,\ \eta = 0$

(b) $\rho = 0.95,\ \eta = 0$

(c) $\rho = 0.20,\ \eta = 1$

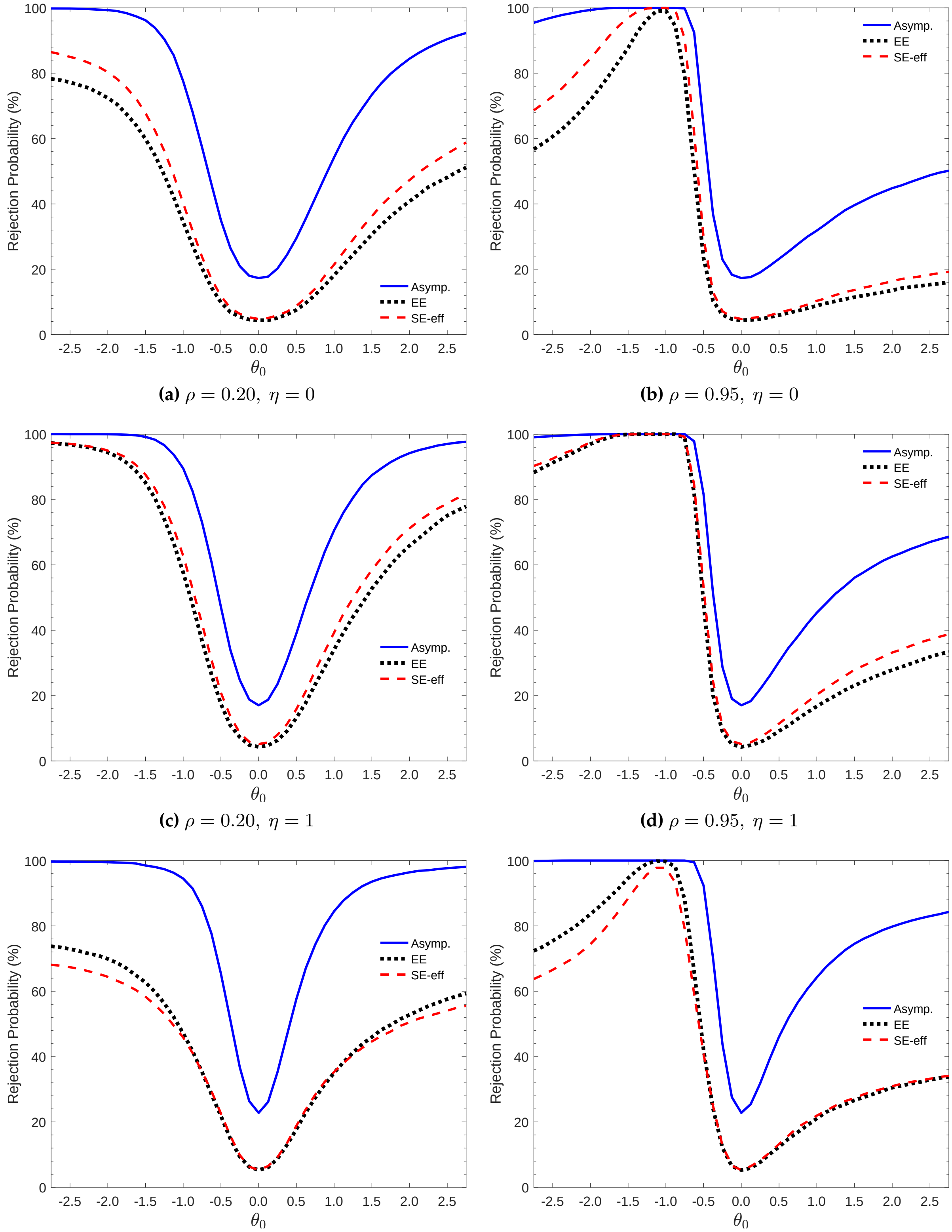
(d) $\rho = 0.95,\ \eta = 1$

(e) $\rho = 0.20,\ \eta = 2$

(f) $\rho = 0.95,\ \eta = 2$

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 999 bootstrap replications for each simulation using Rademacher weights and a DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)\left(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g}\right)^{\kappa}$ evaluated at $\kappa = 1$.

Figure S.3: Power comparisons of bootstrapped AR tests with noncentrality parameter of first-stage F-test at $\mu_{k_z} = 18$ and residual heterogeneity at $\kappa = 2$.



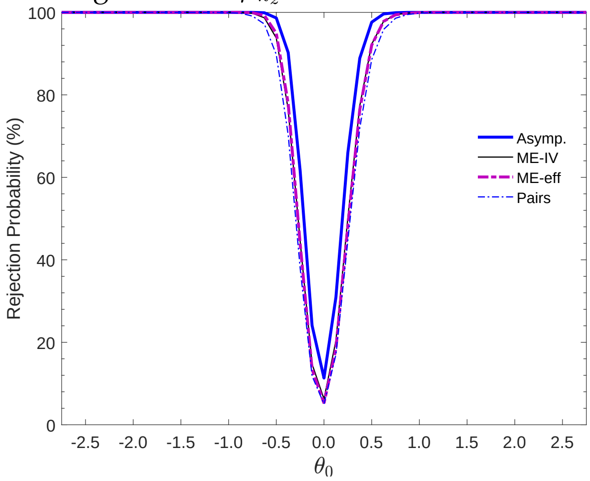
(a) $\rho = 0.20,\ \eta = 0$ **(b)** $\rho = 0.95,\ \eta = 0$

(c) $\rho = 0.20,\ \eta = 1$ **(d)** $\rho = 0.95,\ \eta = 1$

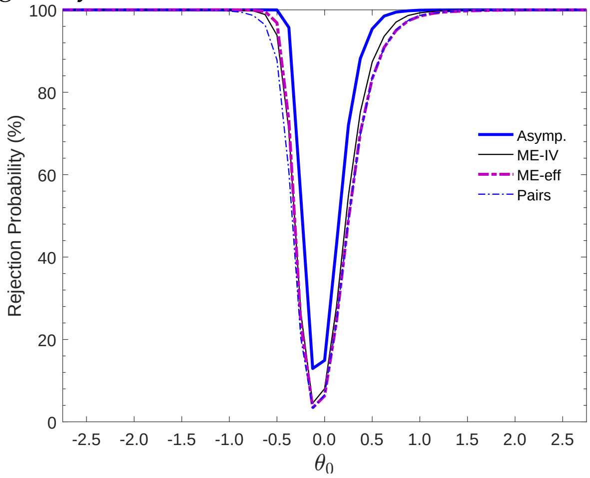
(e) $\rho = 0.20,\ \eta = 2$ **(f)** $\rho = 0.95,\ \eta = 2$

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 999 bootstrap replications for each simulation using Rademacher weights. DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ means 20 observations per cluster, $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$ evaluated at $\kappa = 2$.

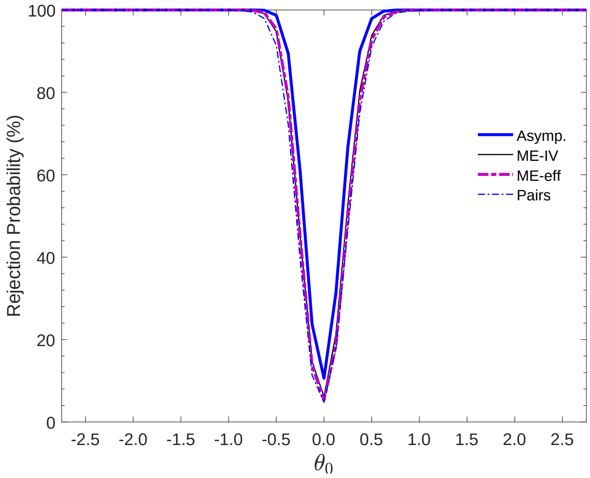
Figure S.4: Power comparisons of bootstrapped Wald tests with noncentrality parameter of first-stage F-test at $\mu_{k_z} = 18$ and residual heterogeneity at $\kappa = 1$.



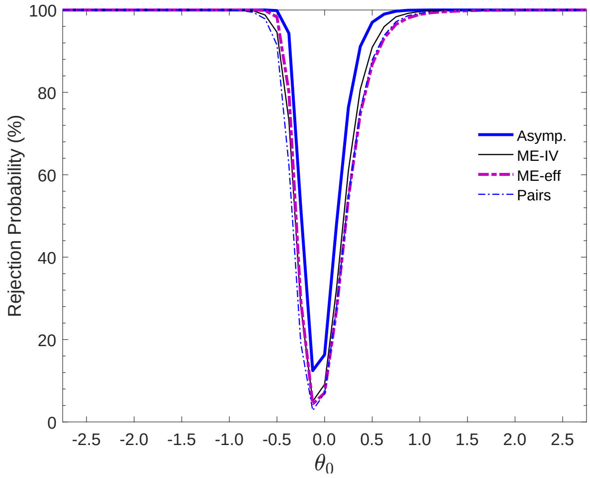
(a) $\rho = 0.20, \eta = 0$



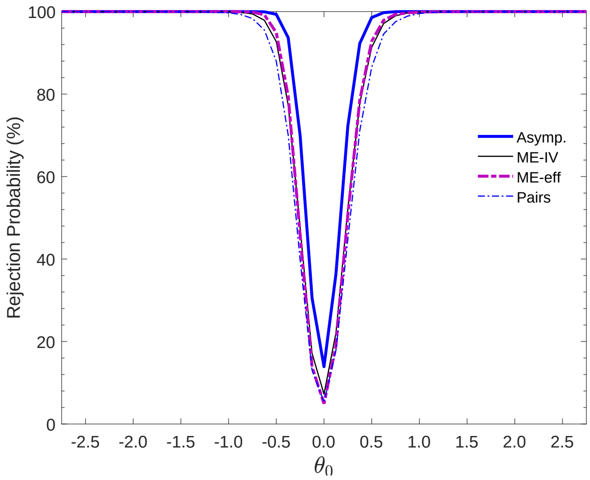
(b) $\rho = 0.95, \eta = 0$



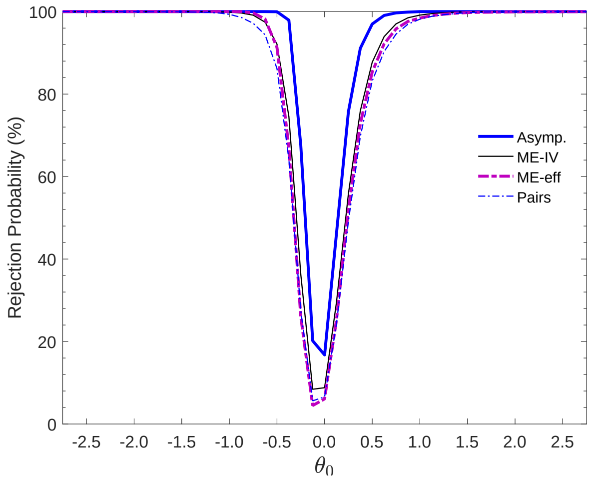
(c) $\rho = 0.20, \eta = 1$



(d) $\rho = 0.95, \eta = 1$



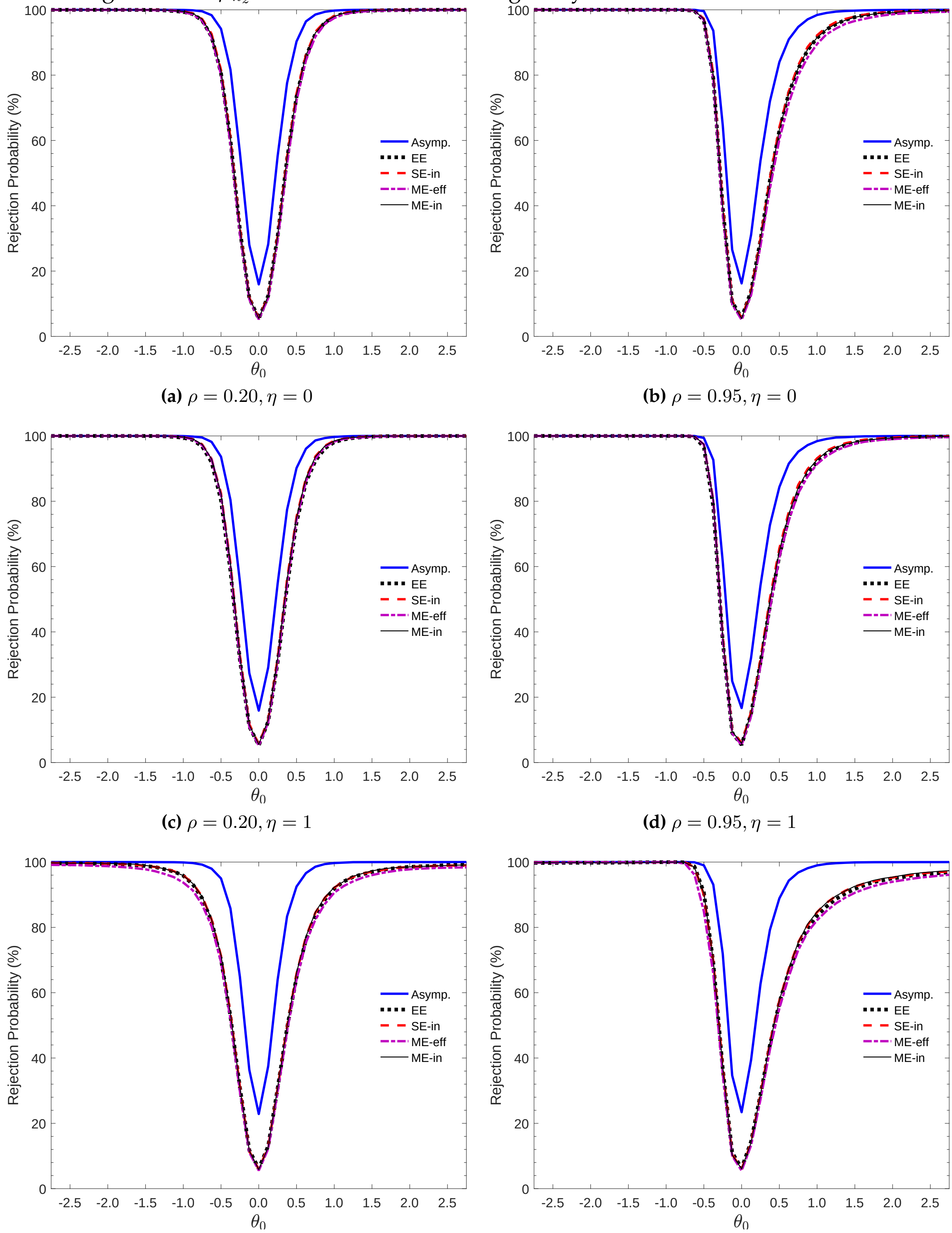
(e) $\rho = 0.20, \eta = 2$



(f) $\rho = 0.95, \eta = 2$

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 999 bootstrap replications for each simulation using Rademacher weights and a DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5 \,/\, k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$ evaluated at $\kappa = 1$.

Figure S.5: Power comparisons of bootstrapped KLM statistics with noncentrality parameter of first-stage F-test at $\mu_{k_z} = 18$ and residual heterogeneity at $\kappa = 1$.



(a) $\rho = 0.20, \eta = 0$

(b) $\rho = 0.95, \eta = 0$

(c) $\rho = 0.20, \eta = 1$

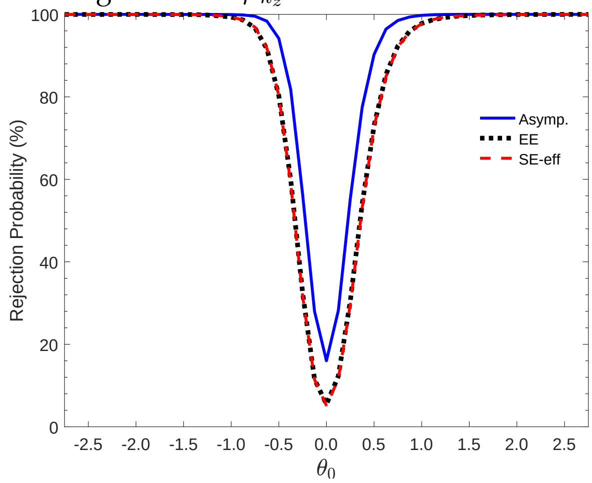
(d) $\rho = 0.95, \eta = 1$

(e) $\rho = 0.20, \eta = 2$

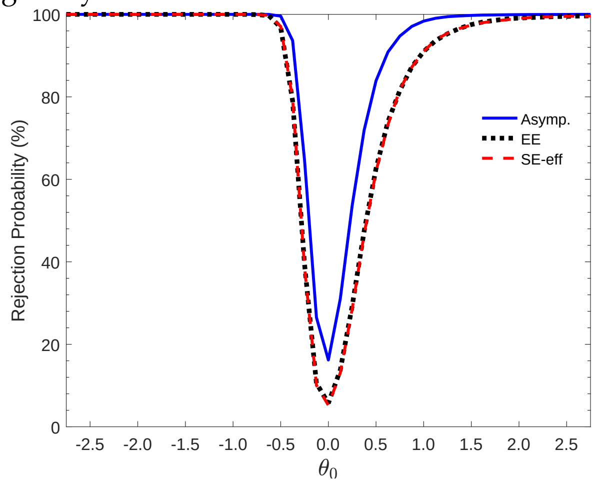
(f) $\rho = 0.95, \eta = 2$

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 999 bootstrap replications for each simulation using Rademacher weights, and a DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ indicates 20 observations per cluster, and $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$ evaluated at $\kappa = 1$.

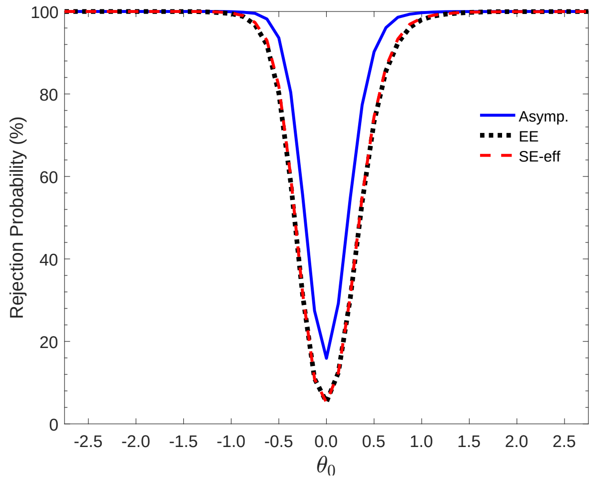
Figure S.6: Power comparisons of bootstrapped CLR statistics with noncentrality parameter of first-stage F-test at $\mu_{k_z} = 18$ and residual heterogeneity at $\kappa = 1$.



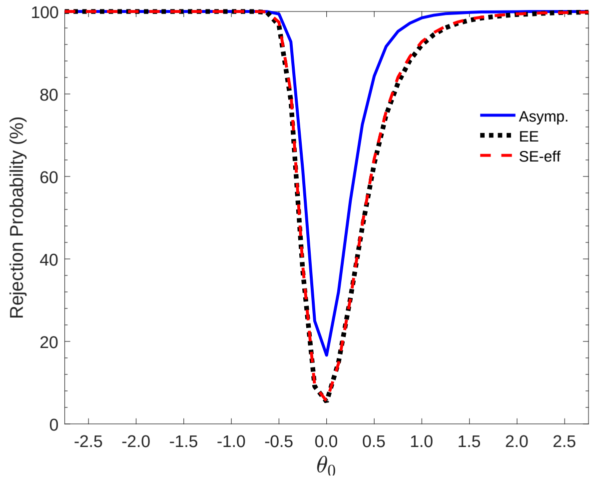
(a) $\rho = 0.20, \eta = 0$



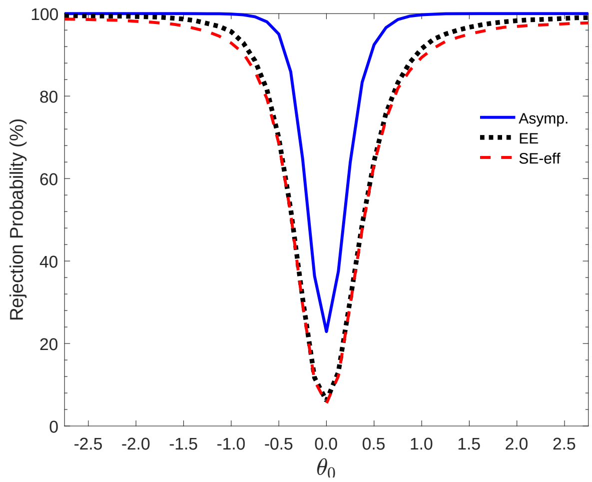
(b) $\rho = 0.95, \eta = 0$



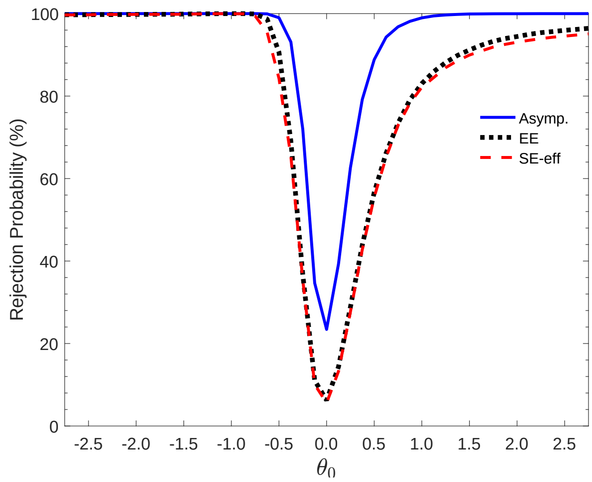
(c) $\rho = 0.20, \eta = 1$



(d) $\rho = 0.95, \eta = 1$



(e) $\rho = 0.20, \eta = 2$



(f) $\rho = 0.95, \eta = 2$

Notes: Authors' calculation from 10,000 Monte Carlo simulations with 999 bootstrap replications for each simulation using Rademacher weights. DGP with normal errors. The sample size is 400 observations with 20 clusters: $\eta = 0$ means 20 observations per cluster, $\eta = 1$ and $\eta = 2$ indicate observations per cluster in the range of 12 to 29 and 7 to 42, respectively. Number of excluded/included instruments: $k_z = 5$ / $k_x = 1$. Within cluster error correlation: $\phi = 0.5$. Skedastic function: $f(\mathbf{z}_{1,g}, \kappa) = h(\kappa)(\boldsymbol{\iota}_g + 2\mathbf{z}_{1,g})^{\kappa}$ evaluated at $\kappa = 1$.